DOCUMENT RESUME

ED 054 202 TM 000 773

TITLE	Testing in Turmoil: A Conference on Problems and Issues in Educational Measurement.
INSTITUTION	Educational Records Bureau, Greenwich, Conn.
PUB DATE	Oct 70
NOTE	46p.; Paper presented at the Thirty-Fifth Annual Meeting of the Educational Records Bureau, New York, New York, October 29-30, 1970
EDRS PRICE	EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS	Achievement Tests, Admission Criteria, *Conference Reports, *Criterion Referenced Tests, Cultural Factors, Decision Making, Educational Accountability, Educational Improvement, Educational Needs, Individualized Instruction, *Measurement, Measurement Goals, Measurement Techniques, *Reading Tests, Testing, *Testing Problems, Testing Programs, Test Interpretation, Test Results

ABSTRACT

The 1970 Educational Conference sponsored by the Educational Records Bureau focused on the topic "Testing in Turmoil: A Conference on Problems and Issues in Educational Measurement." The International Reading Association and the National Council on Measurement in Education co-sponsored two conference sessions entitled "The Measurement of Reading: Procedures and Problems," and "Criterion-Referenced Measures: Pros and Cons" respectively. (CK)

ED 054202

EDUCATIONAL RECORDS BUREAU ERB

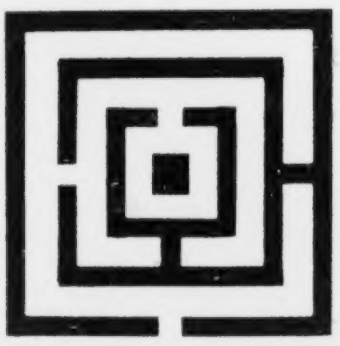
116 MAPLE AVE GREENWICH CONN 06830

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

TM 000 273

Testing in Turmoil:
A Conference on Problems and Issues In Educational Measurement

THE THIRTY-FIFTH ANNUAL CONFERENCE OF EDUCATIONAL RECORDS BUREAU



Hotel Roosevelt **New York, N.Y.**

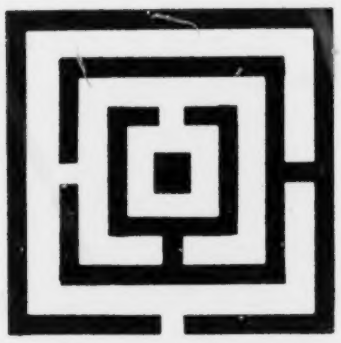
October 29-30, 1970

EDUCATIONAL RECORDS BUREAU ERB

116 MAPLE AVE GREENWICH CONN 06830

Testing in Turmoil:
A Conference on Problems and Issues In Educational Measurement

THE THIRTY-FIFTH ANNUAL CONFERENCE OF EDUCATIONAL RECORDS BUREAU



Hotel Roosevelt **New York, N.Y.**

October 29-30, 1970

Reported and printed by
Thyra D. Ellis and Associates
500 9th Avenue North
Jacksonville Beach, Florida 32250
(904) 246-3380

FOREWORD

The 1970 Educational Conference was the thirty-fifth annual meeting sponsored by Educational Records Bureau. National concern about the assessment of education, and the dispute on how testing should be incorporated into such an assessment, gave the topic special significance. "Testing in Turmoil: A Conference on Problems and Issues in Educational Measurement" proved to be a fitting title as speakers and guests presented a wide range of opinion and fact regarding the use of tests. A careful reading of these proceedings will provide insight into current thought being applied to the problems of educational measurement.

Through the cooperative efforts of the International Reading Association and the National Council on Measurement in Education, two sessions co-sponsored with Educational Records Bureau were designed to fit within the general theme. The IRA topic, "The Measurement of Reading: Procedures and Problems," provided a penetrating discussion of the reading process. The topic also covered the way in which testing is used and misused in the assessment of student achievement in reading. The panel from the National Council on Measurement in Education discussed the timely subject titled "Criterion-Referenced Measures: Pros and Cons." Presenters pointed out that this is not a new movement, but rather a new presentation of an older concept. Advanced technology and improved techniques made the entire issue worthy of reconsideration.

The Board of Trustees and executive staff of Educational Records Bureau is deeply indebted to the many speakers and panelists who helped to make this conference one of the most exciting ones ERB has held in years. The enthusiastic response of the great majority of persons in attendance at the conference pays significant tribute to the efforts put forth by each program participant. In addition, one cannot overlook the contributions make by the staff members of ERB. Their high-spirited enthusiasm during the planning stage and dedication to the difficult task of preparing for this conference truly inspired me in my first involvement with an Educational Records Bureau Conference.

James L. Angel
President

TABLE OF CONTENTS

THURSDAY MORNING SESSION, October 29, 1970

Call to Order and Welcome,
Chairman Edward M. Friedlander..... 1
Introduction of Keynote Speaker,
James H. McKee Quinn..... 1
Keynote Speaker, Lawrence A. Appley
Address: "Measurement - Another Victim of Anti-Excellence"..... 1

THURSDAY MORNING SESSION, SESSION TWO, October 29, 1970

Annual Business Session of Member Schools
Educational Records Bureau
James H. McKee Quinn, Chairman
James L. Angel, President..... 4
Tribute to John Lester, Jr.,
Read by Hart Fessenden
Prepared by Ben D. Wood 4
Report of Combined Meeting of ERB Committees..... 5

THURSDAY LUNCHEON SESSION, October 29, 1970

Introduction of Speaker,
Chairman David D. Hume..... 5
Luncheon Speaker, Manford Byrd, Jr.
Address: "Testing Under Fire: Chicago's Problem"..... 5

THURSDAY AFTERNOON SESSION, SESSION ONE, October 29, 1970

Session: "Building A School Testing Program"
Session Chairman, F. Martin Brown
Program Chairman, Frank B. Womer
Panelist, Jean P. Carten
Panelist, Donald Roberts
Panelist, Daniel Wagner..... 8

THURSDAY AFTERNOON SESSION, SESSION TWO, October 29, 1970

Session: "Ethnic and Cultural Issues in Measurement"
Session Chairman, Wellington V. Grimes
Program Chairman, Richard C. Kelsey
Panelist, Paul Collins
Panelist, Charles Hicks
Panelist, Mrs. Joyce Hicks 10

THURSDAY EVENING SESSION, SESSION ONE, October 29, 1970

Session: "Admissions and Admissions Testing"
Session Chairman, Margaret T. Corey
Program Chairman, Walter M. Birge
Panelist, The Reverend Canon Harold R. Landon
Panelist, Paul G. Sanderson, Jr. 16

THURSDAY EVENING SESSION, SESSION TWO, October 29, 1970

Session: "Interpreting Test Results"
Chairman and Presenter, Harry J. Clawar
Presenter, Edward M. Friedlander..... 17

FRIDAY MORNING SESSION, October 30, 1970

Session: "Educational Accountability and the Measurement Task"
Session Chairman, Richard A. Schlegel
Program Chairman, Robert E. Stake
Panelist, Donald G. Emery
Panelist, E. Gary Joselyn
Panelist, George H. Stern 18

FRIDAY MORNING SESSION, SESSION TWO, October 30, 1970

Session: "Individualized Instruction: The Measurement Dilemma"
Chairman, Harry K. Herrick
Speaker, Uvaldo Palomares.....23

FRIDAY LUNCHEON SESSION, October 30, 1970

Introduction of Speaker, William W. Turnbull
Luncheon Speaker, Jose M. R. Delgado, M.D. 28

FRIDAY AFTERNOON SESSION, SESSION ONE, October 30, 1970

International Reading Association
Session: "The Measurement of Reading: Procedures and Problems"
Chairman, Ralph Staiger
Presenter, Roger Farr, Address: "Testing and Decision Making" 30
Presenter, Walter M. MacGinitie, Address: "What Are We Testing in Reading" 33

FRIDAY AFTERNOON SESSION, SESSION TWO, October 30, 1970

National Council on Measurement In Education
Session: "Criterion-Referenced Measures: Pros and Cons"
Chairman, Elizabeth L. Hagen
Presenter, Robert L. Ebel, Address: "Some Limitations of Criterion-Referenced Measurement"..... 35
Presenter, Anthony J. Nitko, Address: "Criterion-Referenced Testing in the Context of Introduction"..... 37
Discussant, Frederick B. Davis, Address: "Criterion-Referenced Tests" 40

EDUCATIONAL RECORDS BUREAU

Thirty-Fifth Annual Conference

Hotel Roosevelt New York, New York

October 29-30, 1970

- - -

THURSDAY MORNING SESSION

October 29, 1970

- - -

The Thirty-Fifth Annual Convention of the Educational Records Bureau convened in the Grand Ballroom of the Hotel Roosevelt, New York, N. Y., Thursday morning, October 29, 1970, and was called to order at 10:05 o'clock a.m. by Mr. Edward M. Friedlander, Director of the Division of Measurement and Consulting Services.

MR. EDWARD M. FRIEDLANDER: I would like to welcome you this morning to the 35th annual conference. We are looking forward to a very successful and enjoyable time in these two days.

Now, I would like to introduce the Chairman of the Board of Trustees of the Educational Records Bureau, who is also the Headmaster of the Episcopal Academy in Philadelphia, Pa., Mr. James H. McKee Quinn.

CHAIRMAN JAMES H. McKEE QUINN: It is a great pleasure to welcome all of you to our 35th annual conference. I know your attendance in such large numbers is a tribute to the quality of the program. So, I hope you will enjoy it and I know you will.

The quality begins right at the start. It is a privilege for us to have as our keynote speaker, Mr. Lawrence A. Appley, who is Chairman of the Board and former president of the American Management Association. Mr. Appley is a graduate of Ohio Wesleyan University and for a while we almost had him in the educational world. He taught at Colgate University. He holds honorary doctorates from four institutions. Unfortunately he was lured away from education -- for a while at least -- and during his business career he has been associated with the Mobile Oil Company, Richardson-Merrill, and Montgomery Ward, in positions of increasing importance.

During the Second World War, Mr. Appley served as Assistant Secretary of War and later as Executive Director and Deputy Chairman of the War Manpower Commission. In 1944 he was awarded, by the War Department, a citation for Meritorious Civilian Service; and in 1946 he was awarded the Medal for Merit by the President of the United States. He is also a recipient of the 1963 Henry Laurence Gantt Medal. He is a Fellow of the International Academy of Management and has authored four books on the subject of management. Mr. Appley maintains his interest in education. He is a trustee of the Northfield and Mount Herman Schools and Colgate University; and a director of 11 corporations.

We are honored, sir, to have you as the keynote speaker for this conference. Mr. Appley will speak on "Measurement -- Another Victim of Anti-Excellence." Following his talk, Mr. Appley has very graciously agreed to answer questions.

MR. LAWRENCE A. APPLEY: While this message has been put in writing, it makes no claim to being a learned paper. It is an expression from the heart and soul of a man who has been in the educational business for over 40 years -- first as a one-room district school teacher, next as a high school teacher, then as a college instructor, later in education for adult managers, and, in addition to all this, as a practitioner of management. The writings, thoughts, studies, and analyses of many other people have become blended into my own thinking. I claim nothing as original, but neither can I take apart into pieces what I know and think, and give footnotes.

It may be pure coincidence or it may be psychological that history seems to divide itself into decades. We speak of the Roaring '20s, the Depression '30s, the Warring '40s, the Dynamic '50s, and the Booming '60s. How will the '70s and the '80s be characterized? First let me give you my impression of the last 20 years and then make some predictions as to the next 20.

In my humble opinion, the last 20 years can be called the greatest period of dehumanization in the history of this great nation of ours. These last two decades are now in history for their fantastic and unbelievable technological and economic advancement and development. Man has been so busy acquiring a high standard of living and getting to the moon that he has not given enough attention to using all these developments for the good of man rather than for his destruction. Man has become a number on a computer card.

Look at our technological development in the last 20 years. First, consider what has happened to the speed of man. He started out on his stomach. Then he crawled. Then he got on his hands and knees. Then he stood up and walked; then ran; got on the back of an animal; cut down a tree, took the heart out of it and put a sail on it. Then he ultimately put steam in it. In the year 1950, after all the millions of years man had existed, he attained the fantastic speed of 740 miles per hour. He now travels at 27,000 miles an hour!

Consider the explosive power make available to man. He started out with his bare fists. Then he got a slingshot, a catapult, and ultimately gunpowder. In the year 1800, the standard measurement for explosive power was two pounds of black gunpowder. In 1950, it was one ton of TNT. Now, it is one megaton -- one million tons -- of TNT. We can put more explosive power on one B52 than was fired in all of World War II by both sides.

Ninety percent of all the scientists who ever lived are still alive. If you wanted to keep up with the development of technological knowledge, you would have to take a four-year college course every seven years. A college sophomore now has to know more about the nucleus than Niels Bohr knew about it when he got the Nobel Prize for it 40 years ago. This is what is meant when we say that economic and technical growth has exceeded man's capacity to use such growth for the benefit of man rather than for his destruction.

As a result, the human being is rebelling. Murder and crime on our streets, constant wars and threats of wars between nations, poverty in our midst, social and racial strife, and campus revolutions are products of a fantastic leadership vacuum. Where are the great statesmen of today? Name them quickly in your own minds from the areas of religion, education, government, business, labor, et ceteras. Statesmanship is leadership, and leadership is contribution to human development.

The basic institutions where human development starts and is nurtured are: the Home, the Church, and the School. These three institutions have the specific responsibility of preparing young people to meet the problems of life. This means the development of the human body, the human mind, and the human soul. The "community," including governments, is the environment within which these responsible forces work, but the environment does not have the specific responsibility for development.

Let's take a look at these institutions for human development. First, the home. Does one have to stand here at this time and document the failure of the home as an institution of human development? Mother works, Dad works, and the children are turned over to television as a babysitter. I believe it was the Jesuits who said, hundreds of years ago, "Give me the child until he is seven, and I'll give you the man." What kind of seven-year-olds are coming out of our homes today? What has happened to gracious living, the family meal, the evenings together? Where is the discipline, out of which came character and responsibility? A home is not a matter of economic level. It is a matter of parental capability to start the development of children who will grow to be finer and better than the parents.

If anyone needs evidence as to reduced sales and reading of the Bible; reduced church membership, attendance, and participation; reduced financial support of religious institutions, there is plenty of it available. My files are full of it. I can see no particular reason, however, to catalog it in this paper. It would seem that all I need to do is to point out how the authority of the Pope is now being challenged. Nothing like it has ever occurred in history. All I should have to mention is the extent to which the Catholic and Protestant Churches are suffering from the drastic declining entrance by young men and women into the priesthood and the ministry.

The Church, in my humble opinion, lacks the leadership and good management required to fulfill its mission, obtain its objectives and lead us to the powerful influences for which it has a responsibility. There are arguments as to what the Church is. I am speaking of the Church as an establishment -- an organization of people who are gathered together to give strength and help to each other in their spiritual development.

A home that does not experience any impact from the Church cannot be a home. Dropping the kids off for Sunday School while Mother and Dad go off to play golf isn't the kind of impact I am talking about. A family has to grow up together with moral standards, ethics, and guidelines. They have to become built in, and spiritual influence is required for that. We are here today, however, to take a look at our educational systems.

Again there is much evidence, there are many testimonials from educators themselves, as to the breakdown in our public and private educational institutions. Education is change, and that change takes place through increasing information and knowledge and through the impact of teachers upon character.

Part of the information imparted is the basic ability to read and write. These are the bases of what we call literacy. They

are proven requirements to human development. The U. S. Commissioner of Education said a little more than a year ago that the capacity of our young people along these lines has been consistently declining.

I happen to be closely identified as a Director or Trustee with five different educational institutions. I have been, and am, working with several others. There are many documents of one kind or another placed in my hands that have been written in longhand by college students and even graduate students. I am appalled at what I read, if I can read it at all. Here is one example from a college graduate and Peace Corps teacher -- "This responsibility is primarily toward people with who you are working." Try another, "The program should give me the skills to do a job good and to do it successful." Still another. "Since being here I have received a general drift of what is to be accomplished and feel that there will be changes within myself and hope to accomplish them."

What do you think of this from a 29-year-old college graduate, now in graduate school? "My reason for being here is to improve my self-concept -- to gain confidence in myself so that I can go out in the business world and be production." Still another, "It's management's responsibility to instill the best effort from everyone."

Undoubtedly listeners and readers will react with "Those are examples of carelessness and bad grammar." Right! As far as I am concerned, however, that is just another way of saying our young people can't read or write. It is fairly well known, I guess, among professional "testers" that there is a direct relationship between the breadth of one's vocabulary and the advancement within the major professions. There is also a relationship between carelessness and reliability on a job.

A two-year study was made by Jean Predseil, Director of the Centre de Recherches et D'Etudes Des Chefs D'Entreprise in Paris, France, in nine different countries on what has been happening on college campuses. The study was called "The Questioning of Education by Youth." Following is a quotation from that study: "Most of the declarations expressing the refusals of university youth and almost all the articles and studies on the subject emphasize the fact that these are not demands concerning the living conditions or the organization, but rather demands putting in question the entire basis of the educational content, of its spirit, of the relations with the professors, of the monopoly of the latter concerning the choice of the programs and their judgment ..... It is a sort of accusation of the previous generation, judged incapable of truly preparing, with the desired effectiveness, today's youth for today's world." If this is what our students are saying, then I agree with them.

Very rarely do you hear "profit" referred to or taught any place in our educational system. It's come to be a dirty word. I believe that is so because those who make it honestly have to have superior skill to do so. "Inspection" is not acceptable to modern workers because it allows for the possibility of poor work. "Standards" imply differences in performance. "Measurement" exposes mediocrity. This is an age that can be characterized as the incompetent in revolt against the competent.

Because of seniority, tenure, and the tendency to avoid standards, measurements, and discipline, the gap between educators and the business world that supports them is becoming wider and wider. While tenure and seniority are products of a system that made them necessary for human protection, they are now outmoded. Society no longer permits the biases and indiscretions that tenure protected our teachers against. Furthermore, there are teachers' organizations that are effectively representative.

As said earlier in this presentation, education is change. If there has been no change, there has been no education. Change can be measured, and the measurement of it indicates the effectiveness of those who are endeavoring to bring about change.

Human development is change in the direction of excellence. Excellence is defined by standards. Measurement determines progress toward the attainment of those standards.

In this era of glorification of mediocrity, in this period of admiring and publicizing anti-excellence, we are faced with a tremendous challenge in the Home, in the Church, and in the School. As you might guess, I have an answer. I wouldn't be here before you if I didn't think it to be an effective answer.

The next 20 years, in my opinion, will be characterized as the greatest years of Humanization (human development) that this nation and possibly the world has ever seen. Human development is management. That is the purpose of management. Better management is demanded right now, and those who meet that demand will be rewarded with a sense of attainment, and those who do not will wither by the wayside.

What's meant by good management? My answer is divided into three parts: the Nature of Management, the Processes of Management, and the Character in Management.

Nature of Management

Management is applicable and needed in any situation where groups of human beings are gathered together in the attainment of a common objective. It is not the exclusive property of business. There is management in religion, education, government, in business, labor unions, on the farm, and in the home. The same principles apply, even though the application of those principles varies.

A manager must know, deep down in his soul, and have burned into his conscience, that management is the development of people so that they may be more effective workers and citizens. Managers must feel to the very tips of their fingers and toes that they are supposed to make things happen. They make the future, they do not wait for it. Management is coaching, it is teaching, it is guiding and motivating.

The Management Processes

Those who expect to manage effectively in the next 20 years will have to be formally trained in the basic processes and in the use of the tools. These apply to the field of education as well as to any other activity.

Management Processes are: taking an inventory of current positions, assets, and liabilities; planning the future; organizing human resources; organizing physical resources; establishing goals and standards; measuring results against the goals and the standards; determining constructive action to be taken to attain excellence; and providing and motivating financial and non-financial rewards and incentives. Each of these processes requires a particular skill. Each requires the mastery of certain tools.

Character in Management

The third requirement of good management is that there be character in it -- real character in it. This means that there is a record of successful attainment and gratification from being of some service to society; that consultative supervision is a way of life which means that those under one's supervision have a great deal of creativity, knowledge, and ideas to contribute to management; that there is a contagious inspirational mission that goes far beyond the making of the almighty buck, it goes beyond selfish interest. In terms of the educator it means that the subject matter that he teaches is merely the medium through which he reaches the character and life of the student. (This latter statement was made by Woodrow Wilson when he was President of Princeton.) There is a basic philosophy as to the existence of a Supreme Being and of a basic plan for civilization; there is emotional stability, which means that there is not much of a gap between one's basic philosophy, what one believes, and the way life makes him live.

These things cannot be developed by chance. They cannot be inherited or acquired from others. They are the result of hours, days, weeks, months and years of intensive, dedicated training.

It is most gratifying to participate in programs wherein universities and colleges are giving intensive time and thought to the development of what they are and where they want to go, and how they are going to get there. This is the process of scientific planning. I have participated in a university program where teachers are deeply involved in the determination of how to measure their effectiveness. What is the difference between a situation where there is a teacher in the classroom and a situation where there is no teacher in the classroom? The answer to that question becomes standards of excellence, and measurements can be made against the attainment of such excellence.

The American Management Association is now working with a large grant from the Federal Office of Education in trying to develop the application of management principles to the public school system in the states of Maryland and North Carolina. This is the result of a very successful experience with the public school system of the City of Syracuse, New York. Within the last few weeks I met with a number of legislators, budget and educational officials of the State of California. Seven of their State College Presidents went through an intensive program of management training, and another nine to twelve of them will be in a similar program within the next month. Consideration is being given to the inclusion in the budget of a line item on management development.

If this is obtained, it will be reflected throughout the entire public school system of the State of California.

What I am talking about is not a hope, a dream, or a mass of theory. Over the past many years a very specific discipline of management has been developed. It can be taught, it can be transmitted. It cannot, however, be acquired out of the atmosphere by the process of osmosis. It has now become a necessity and is no longer a matter of choice. Survival depends upon the acceptance and practice of it.

The time is rapidly coming to a close when large numbers of

our citizenry in this country can avoid responsibility, back away from highly disciplined education and training, and drag other people down to their level of mediocrity. The people of this country will not stand for it. Missions, objectives, and goals must be developed scientifically. Standards of excellence must be established. Individual performance must be measured against those standards of excellence. Intensive training must be provided to bring performance nearer and nearer to standards of excellence. All this must be done as a result of a driving motivation to be proud of one's life, to be proud of one's attainment, to have a sense of value and importance in this world. After all, this is the greatest source of happiness. This is the basic purpose of the plan of civilization.

Idealistic? Yes, 30 years ago, but not today. Impractical? Yes, 30 years ago, but not today. Expensive in time, effort, and money? Yes, and much more so today than 30 years ago.

The day of the amateur in management is past. We are in the age of bigness -- big religious institutions, big educational institutions, big business, big labor, big government, and nothing is going to get smaller; it is all going to get bigger. This demands new thinking, new concepts, new organization structures, new drive, new inspiration in order to bring the individual back into his rightful position of supremacy. Humanization is the order of the day, and that takes good management, which in turn takes intensive training.

CHAIRMAN QUINN: Are there any questions anyone would like to address to Mr. Appley? Would you use the microphones, please.

MEMBER: Is there a copy of Dr. Appley's address available?

CHAIRMAN QUINN: Dr. Appley's address will be printed in the proceedings of the conference which will be available as soon as they are printed.

MEMBER: I have an uneasy feeling you are blaming the patient for his illness. A youngster who doesn't learn to read at four or five or six and who can't express himself fluently at 29 is the product of our educational system and is not necessarily the product of his own dereliction.

I have been fooling around with tests now for 30 years and one of the problems we face is that we can neither define emotional stability nor predict the kind of person that develops from the environments we arrange so nicely.

DR. APPLEY: I agree with the first part of your observation that the child is the product of the system. I meant to imply that. As far as the testing is concerned, I personally do not believe there are very many tests that can tell us what a child will be but many tests tell us what he is and tell us what his aptitudes are; they can't tell us which way he is going to jump.

These are not yet nice, tidy little bundles, but let me suggest this. My belief in testing is not in the validity of the test itself but in the process. And the fact that we use a test means that it causes us to give more attention to the child than we otherwise might give, but I would certainly not want to have my judgment based entirely on test results.

MEMBER: I think we are inclined to assume that there are certain people who can achieve excellence. And I think we have to approach the problem from the other direction. Every individual is capable of achieving a degree of excellence and we have to find the motivation or vehicle in order to exploit his opportunity or potential to be excellent.

DR. APPLEY: I agree again. Only, to further your thought: when Jack Nicklaus became a young man big enough to carry a golf club he proved to be a natural born golfer and his degree of excellence is about as high as you can get in the golf world today. The first occasion I ever showed any interest in golf indicated I never would amount to much, but we both took lessons. Whenever he plays he usually is playing subpar gold. That is a high degree of excellence. When I play I am very lucky if I can get in the 90's. The point is we are both better because we took training. My level of excellence is much lower than his.

I think this is what you are saying (am I right?): You can't raise all children to the same standard. This is why each has to be trained in relation to his own individual profile and this goes for all areas of human development.

MEMBER: I was intrigued by what I detect to be an omission in your example of good grammar mentality. That is the idea that a student wouldn't know whether an adverb or an adjective would follow a copulative verb. That is what I find critical. It reminds me of the president of the Council of English reading 8,000 themes, all with correct capitalization, punctuation and spelling, but he did not find a single idea in the 8,000. And I think this is what we should be concerned with. If the 29-year-old can articulate his concerns to his kids, to his business associates, that is what I think education is all about.

DR. APPLEY: I agree with what you think education is all about, but I still believe in doing what you are saying and doing it well. I just cannot accept carelessness. I think it is all right for one's individuality to blossom and expand and grow. This is what we want. But I think along the way there are certain standards of how one lives with others, how one communicates, that we should not be careless about and I do not think it is one versus the other.

I was involved a few weeks ago in a discussion with faculty, students and trustees. (We are getting more and more into this practice, getting the three groups together.) The question arose as to whether the campus is an instrument of social change or a place where people develop to become instruments of social change. This is a very acute question, and my reference to reading, writing, and arithmetic and the examples I used were to dramatize the point, but I certainly would not wish to overemphasize good spelling.

I can't spell myself. But I believe the basis of our system of communication, consisting of reading, writing, and arithmetic -- the fundamentals, if you wish -- should be learned and then the rest should be left to the initiative and competence of the individual for selective training.

MEMBER: I suggest if you pose the wrong question you get the wrong answer. The question is not whether you want "good grammar or good taste." The answer is we need more of both and we can have both if we want it, and society has to have it. We don't post a dichotomy of choice when choice isn't the proposition.

DR. APPLEY: Is that not approximately what I was trying to say? I think it is! But let me say at the same time that I think it is an insult to the public to brag about bad grammar. I don't think they have to sell their cigarettes that way.

MEMBER: I am not sure I can phrase this question easily, but I presume, Dr. Appley, that you do have some model, some conceptual model in your mind of what constitutes excellence in society which will then establish the standards of excellence which we are about to develop for our human beings. Our trouble seems to be that we develop them, and the society which is not lacking in excellence can no longer absorb the excellent people that we have.

DR. APPLEY: There are several points you are making, but let me see if I can make this brief. My concept of a standard of excellence is this -- and it is in the paper but I didn't take the time to bring it out and should have done so. I am a great believer in the process of consultative supervision.

I, therefore, believe a standard should be developed by the people who are to attain it and I want the standard developed in consultation between the teacher and the student, between the manager and the worker, between the labor union leader and the labor union worker. I want them to develop their own standards of excellence and then try to attain them.

There is no uniform standard of excellence for everybody. At the same time I believe progress in civilization is a society in which standards get higher and higher. The law is a reflection of public standards and growth in standards. Laws are passed to reflect what people want in a democratic society, and we make progress along the way.

Society says, "Let's raise all society a little higher." One of my hangups is the way in which leaders in the business community rebel against any legislation whatsoever, anything that legislates or restricts. The answer is "No," automatically.

My feeling is that civilization progresses through government regulations. We get together as a community and we say we shall not kill; so, we have a government to see that we do not kill each other. Everybody has agreed to it. It is a disciplinary advancement in the standards of society. The speed laws on the road -- we do not want to injure anybody, but we know we can't go out and drive sensibly unless we are aware of the state trooper. By our own selection, we force this standard upon ourselves.

Civilization provides its own standards and patrols them.

MEMBER: You say then, that excellence is an emergent property of society and institution and not something that you look back upon?

DR. APPLEY: Or impose upon others.

MEMBER: I find much of what you said very congenial to me. Yet I find inconsistencies which bother me. In the first part of your talk you mentioned that our civilization is increasing dramatically exponentially and yet we are faced with decreasing excellence.

my are we increasing so tremendously in technical improvement, technical achievement?

DR. APPLEY: I pointed out the tremendous progress we were making technologically and materially and yet I implied there was anti-excellence. I assume you mean how can you reach the moon if you are not pretty excellent?

MEMBER: Right.

DR. APPLEY: The inconsistency I will try to eliminate. I am speaking of excellence in the field of human development. There has been excellence in technical development, excellence in material development, but a decline in excellence in human development. Does this help you?

MEMBER: It does except that all of the references you made, as far as I can tell, related to productivity and creativity, and not to self-development or to things pertaining to the self. They related to civilization.

DR. APPLEY: Well, it was my mistake if I didn't make clear that the process that I was describing is a process of human development. If I want to develop myself, I have to know what I am not, what I want to be, whose help I need to help me be that, what physical things I need, what standards I want to attain and I have to be willing to measure myself against them.

Then I have to take development work to see I do attain those standards, then expect some kind of a reward -- that goes for myself or the group in the classroom or the workshop or wherever you happen to be working.

MEMBER: Most people here know young people who are intelligent and who are seeking a better way, who are really seeking excellence in their own way which isn't perhaps yours or mine. If they were to hear you today, much of what you say would be meaningless to them or worse.

They would reject. They would say we all know this enormous gap. How do you propose that we bridge this gap between the young people in high schools and colleges, since we all know they would not agree with much that you say? Maybe out of ignorance, I don't know. How would you attempt --

DR. APPLEY: The question here is that most of our fine, young --

MEMBER: Not most -- many.

DR. APPLEY: -- many of our young people, and I will say most of them are fine. In fact, I wouldn't want to be the one to say who isn't.

MEMBER: You wouldn't apply your standards?

DR. APPLEY: Right! The question was that our fine young people, many of them, would not accept what I said today. How are we going to close the gap with them? I wouldn't say to young people exactly what I said today because I am speaking today to the educators, to the "coaches," of which I am one.

I am trying to challenge, to issue a challenge to do a better job. When talking with the students, about 10,000 of them a year, on the campuses of our colleges, high schools and grade schools, I find a tremendous response to the appeal of the same approach that I used here as to one's own human development.

Just recently I gave a commencement address to a large student body. I have to be careful that I don't give it to you again. But I made the statement that the chaos in this world is a challenge to human development. And rather than becoming despondent and seeing this world of ours as a place in which there is no challenge, we should realize that the world is full of challenge and the plan for civilization is that each generation shall have more difficult problems to solve than the last. That is the way human progress is made.

You never become a better tennis player by always beating your opposition. You never improve at bridge by always winning. Competition must get tougher. With every generation the problems are greater. It is much more difficult to control the use of the Atom Bomb than it was to control the bow and arrow. It is much more difficult to integrate the black man than to segregate him.

These problems are extremely difficult and, therefore, it is going to take fine people to meet these challenges and they must be formally trained in how to do it.

I am afraid my message has some of the flavor of the efficiency expert in it and I really hope that is is not coming through that way. My message is that if we go about human development, self-development, and the development of others in an orderly way, we will be more effective if we do it with a "hit-or-miss," leave-it-to-chance, day-in-and-day-out approach. This is my whole theme.

CHAIRMAN QUINN: I think on that note we will bring the meeting to a close. We are very grateful to you, Dr. Appley, for coming down today.

THURSDAY MORNING SESSION

Second Session

October 29, 1970

- - -

The meeting of the members of the Educational Records Bureau was called to order by Chairman James H. McKee Quinn. Mr. Quinn presented a brief report on the activities of ERB for the year, and announced the results of the election of Board of Trustee members. The following persons were elected to the Board for six-year terms:

INDEPENDENT SCHOOLS:
David Pynchon
Headmaster
Deerfield Academy
Deerfield, Massachusetts

TWO-YEAR COLLEGES:
Mrs. Livingston Hall
Headmistress
Simon's Rock School
Great Barrington, Massachusetts

FOUR-YEAR COLLEGES: Approved Representatives

Harry Coleman, Dean
Columbia College
Columbia University
New York, New York

Nathaniel S. French
Department of Education
University of Massachusetts
Amherst, Massachusetts

Frank B. Womer, Staff Director
National Assessment Program
2222 Fuller Road Apt. 29A
University of Michigan
Ann Arbor, Michigan 48105

Mr. James L. Angel, President of ERB, presented the annual report to members, which was distributed at the meeting and later mailed to all member institutions.

Mr. Hart Fessenden read a tribute to John Lester, Sr., deceased, an original trustee at the time of the founding of Educational Records Bureau. The tribute, prepared by Dr. Ben Wood, follows:

JOHN ASHBY LESTER
August 1, 1871 - September 4, 1969

IN MEMORIAM

On September 4, 1969, Dr. John A. Lester died in Rosemont Manor, Pennsylvania, in his 98th year.

Dr. Lester was one of the most able and dedicated of the founding fathers of the ERB -- that small group of thinkers who, despite opposition, went ahead (with courage equal to their now universally accepted humane insights) and established a testing and educational service organization that has had a large, pioneering part in formulating and disseminating the ideas that have become an essential part of the foundation of the current revolution in the purposes, methods, and implementations of the testing and guidance movement.

Dr. Lester had a large part in spreading understanding and acceptance of the concept of education as learning by individual pupils, each at his individually appropriate level, and at his individually appropriate pace, thus promoting the powerful, multiple advantages of success-motivated study and learning, for moral as well as cognitive goals of education.

For several decades of his scholarly and fruitful professional career Dr. Lester was almost a lone voice crying in the wilderness, explaining how educators might secure these powerful advantages, which are still being thrown into reverse in far too many of our increasingly costly schools and classrooms by routines and practices which many of our most thoughtful educators and writers openly identify as relics of the barbarous aspects of early schools, which Comenius described as "slaughterhouses of the young."

Dr. Lester was far too gentle and kindly to use such harsh adjectives. Instead of cursing the pervasive darkness that blighted so many of our schools, his habit was to try unceasingly to light a candle. It is a great consolation to all of us who mourn his passing that he lived to see many of his candles grow into such flaming lights as are exhibited in the writings of several stars in the new galaxy of educational thinkers.

Dr. Lester was an active member of the Board of Trustees for nearly two decades. In the record of the Conference there will be a tribute to him and a short history of his life; but let us stand now in silent appreciation and grateful memory of a wise educator and a widely beloved colleague who contributed so much to establishing the ERB and guiding its activities to truly benign pruposes.

REPORT OF COMBINED MEETING OF
VARIOUS ERB COMMITTEES

A breakfast meeting was held Friday morning, October 30, 1970, for the members of the four major committees, the Committee on Tests and Measurements; the Independent School Advisory Committee; The Public School Advisory Committee; and the School and College Relations Committee.

Mr. Angel presented the revised program for committee involvement as authorized by the Board of Trustees. All existing committees of the Bureau were phased out to provide for the appointment of "ad hoc" or "task force" committees, where members would be appointed for specific problems to be resolved and only for the length of time the study would be in process. An Advisory Council will be appointed by the Board of Trustees, made up of no more than seven members, which will act as a principal consulting body to the President and the Board of Trustees. Two subcommittees, the Test Selection and Mathematics Subcommittees, will be retained as "task force" committees until current assignments are completed.

Committee members discussed the changes at some length with general agreement expressed that the change in committee involvement should provide more meaningful and dynamic participation by members. Meeting adjourned.

- - -

THURSDAY LUNCHEON SESSION

October 29, 1970

- - -

The Thursday luncheon session of the Educational Records Bureau convened in the Terrace Suite of the Hotel Roosevelt, New York, N. Y., October 29, 1970, at 12:30 p.m. with David D. Hume, Chairman, presiding.

The program following the luncheon began at 1:30 p.m.

CHAIRMAN DAVID D. HUME: When Jim Angel asked me to chair this luncheon meeting today, I wasn't sure how many things I would be up to my neck in at this time of the year. But I said I would do so, and I am pleased to be here.

I will keep these remarks brief. First, I would like to introduce the people sitting here who dignify the head table. On my left is Tony Barber, Headmaster of the Laurence School and a trustee of the organization. Next to him is Mr. Hart Fessenden, Headmaster Emeritus of the Fessenden Schools, also a trustee.

And next to him, Jack Gummere, Headmaster Emeritus of The William Penn Charter School, also a trustee of the organization. We are heavy on Emeriti. Even more distinguished, next we have the most Emeritus of them all, Ben Wood. On my far right is Bob Lynn, Headmaster of the Memphis University School. Next to him is Jim Angel, President of ERB.

Next is Jim Quinn, Headmaster of the Episcopal Academy, Philadelphia and Chairman of the Board of the Educational Records Bureau. Now, I would like to introduce to you our speaker.

He was graduated from Central College in Iowa in 1949; did graduate work in Educational Psychology at the Atlanta University, at DePaul University and the University of Chicago. He took his Master's Degree in 1954 from Atlanta University. He taught in the Quincy, Illinois and Chicago public schools.

There he was a teacher, a master teacher, and a school principal. Since that time he has been Deputy Superintendent of Schools for the City of Chicago. This is an enormous job. He tells me there are 500,000 students and the Deputy Superintendent is the man responsible for the day-to-day operations of that entire system. Mr. Manford Byrd, Jr.

MR. MANFORD BYRD, JR.: Mr. Hume, table guests, ladies and gentlemen; I am indeed happy to be with you during this two-day conference for several reasons. When I was invited to come and share with you some of the problems of the Chicago public schools, I suppose I felt inwardly I should take advantage of this opportunity to talk about my problems with anybody who would like to listen, and I am convinced that it is far better to be talking about the problems than to be on the scene, on the spot, facing them right now. So, I am taking this respite and enjoying it.

When I considered the theme "Testing In Turmoil -- A Conference on Problems and Issues in Educational Measurement," it occurred to me that it should not be a surprise to any of us that testing is now coming in for its share of criticism as part of the turmoil in the educational scene.

Indeed, I have confronted and experienced turmoil in just about every activity I have attempted in the educational field. And these activities range from the construction of modular buildings, teacher assignment programs, and the bussing of students, to the assignment of principles, and sex education.

And interestingly enough, when you have these confrontations, you get some offshoots or fallout that, in spite of it all, strike you as being rather humorous. I recall talking to an irate parent about sex education and during our conference one of the ladies demonstrated to me why I should not invade this realm. This was the sacred realm of the parent and I had better stay out of it. And she said, "I want you to stay out for several reasons, but one of them is this: I am afraid if you handle this problem the way you handle reading the race will become extinct."

I sat and talked with some of my table colleagues and others during lunch and I must admit to you I rather wondered why you invited me, as I notice we have among the guests so many persons from independent or private school systems, and I thought of an anecdote.

This story is told that a businessman on his way home from work met a beggar dressed all in rags, who cried, "Mister, can you spare a quarter?"

And the businessman said, "What do you want with a quarter, sir?"

"I need it."

He said, "Do you drink?"

"No."

"Do you smoke?"

"No."

"Do you gamble?"

"No."

"You come on home with me and I will give you a dollar."

The beggar thought this was a good deal and went along with him. The businessman got to the door, walked in and said to his wife, "Hey, dear, come here, I want to show you something. Here is what a person looks like who doesn't gamble, doesn't drink, and doesn't smoke."

I had this feeling that maybe you asked me to come in so you could take a look at how a big city administrator, embattled with problems, reacts at this time. Nonetheless, I am delighted to be with you and to share with you, for a moment, some of Chicago's concerns and Chicago's problems relative to testing.

The Chicago public schools instituted a citywide testing program in elementary school grades in 1936 and in the high school grades in 1937. From the inception of the standardized testing programs, until about 10 years ago, each school chose its tests from our official list.

We gathered teachers, administrators, and counselors to prepare the testing list but each school was left to its own to select the test to be used. I might add that annually we are testing over a quarter of a million youngsters in the Chicago public schools in citywide programs, to say nothing about the many independent schools.

With the implementation of the National Defense Education Act we began both a long-range processing of the output and a citywide adoption of the tests used with the selection being made from the approved list by comparable committee process. This change took about seven years to complete.

We tested six points -- first, third, sixth, eighth, ninth, and eleventh grades. Actually, in this we worked downward, beginning in the high school grades and going down through the elementary grades. When the sixth and eighth grade

programs were to be converted in 1961 and the citywide committee reviewed the tests on approved lists, and recommendations were to be made, I am told the majority opinion came under fire.

I mention this only to indicate that way back then when things were relatively quiet there was turmoil in testing. The program was reviewed a few years later and is up for review now. The current review committee will be expected to deal with issues and problems and policy recommendations which previous committees felt no need to pursue.

Beginning around 1962 or 1963 pressure began to be put upon the school administration to make test results public. As a result presentations were made at public board meetings on citywide data and later of anonymous schools, but this did not satisfy and the clamor continued. As a matter of fact, when I joined the superintendent's staff in July, 1967, my first assignment by the General Superintendent -- was to devise a means of reporting individual building results to the public.

Last year for the first time we issued median tests scores -- school by school -- for each grade level tested during the previous school year. Each of our over 500 schools had a page to itself and as a result the book is known to some as the "telephone book."

To others it is the "green dragon", or monster, in deference to the color of the cover, but in tune with some of the attitudes for the release and controversy generated by it. The book had a substantial introduction on which the staff spent quite a bit of time in an attempt to put testing in the proper prospective. The staff has worked hard in this introduction and other methods to get our point across.

I have taken this time to review or give an overview of testing in Chicago for a few basic reasons.

You need to know the experience of the school system from which I speak. I think this review pretty well embraces the experience of other school systems, some of which had far more pressure on them than we have. The review carries us from the days when turmoil of testing was non-existent, when it began as an in-house affair. Pressure to release the data continued and new pressure are developing, especially from our Puerto Rican community (more on that later).

I think it is also important to cite the role and development of school standardized testings in Chicago brought about by the National Defense Education Act of 1958. There is no question as to its impetus for testing. NDEA's was, I believe, the first positive response of the Federal Administration to Russia's Sputnik lofted in October 1957. Previous negative responses from the Federal Administration castigated public high schools of the country because we did not get into orbit first.

In this connection, a passage from Arnold Toynbee's "Civilization on Trial" seems pertinent. I think we must all today, in our trouble, take care that this does not apply to us. "It is always a test of character to be baffled and up against it, but the test is particularly severe when the adversity comes suddenly at the noon of the halcyon day, and one expected to endure to eternity."

I skip a sentence and continue. "The act to pass the buck in adversity is still more dangerous than to persuade one's self that prosperity is everlasting." That is the end of the quotation but it is not the end of the idea.

The buck was passed to public education, especially secondary education following Sputnik, and adversity has endured since then. Causes of adversity have, in fact, compounded since then. For the last several years results of school standardized testing have brought schools under fire, with every major city in the country in trouble when its results have been compared with the norm group.

One begins to ask where the problem lies and what is its nature? What are its dimensions? I want to turn now to some of these problems and issues without presuming to exhaust either roster. In fact, in organizing a statement I have found that it is difficult to be sure always which point is a problem and which is an issue.

I will start off with what is a difficult problem, but one which I think is only by inference an issue. That is the problem of lag and of the inability of some ponderous enterprise to keep abreast of changes and realities. Let me begin with an illustration from World War II days involving individuals' psychological testing.

I am told that the Wechsler test on a certain day included as an easy question, a query as to the name of the previous President of the United States, and as a difficult question, "Where is Tokyo?" In the 1940's, after war began with Japan, these questions virtually changed places on the scale of difficulty. Every school child with a brother, uncle or father in the service knew where Tokyo was. Conversely, with Roosevelt in the Presidency since their infancy and early childhood, Hoover was a name unknown to them. Those of us in the business know that test builders study curriculum guides across the country and testbooks, too. We understand that there is a built-in lag, so that it takes several years, for example, for testbooks to carry illustrations of non-white faces, to say nothing of introducing appropriate explanations of ethical and racial contributions to our national life.

We know that test changes follow curriculum textbooks and changes. But this does not pacify our critics, and I do not think it should. There are a dozen or so major cities where most of the confrontations take place through a great many tests, but they do not dominate the companies' sales. Big city constituents just do not understand this.

As a matter of fact, I must admit I do not understand why we do not have a greater influence in the development of these instruments. We know now that test scores can be manipulated to give us any sort of distribution we want because there are definite mathematical variances to be obtained from normal distribution.

One of the aims of present day test builders is the construction of tests that will give normal distributions for the type of population in which they are to be used. The question being raised today that brings the testing program under fire is quite simply: are the school populations of large cities comparable to the norm groups? That is to say, are they adequately represented in the norm process? Enlightened believers in testing are beginning to question not testing or its value but the tests. However, I think the test makers and others are questioning and will question our educational programs continuously.

The problem is compounded by the fact that the public still believes the median represents what every Tom, Dick and Harry not only should but must achieve for a school to be doing its job. I am afraid standardized testing has unwittingly reinforced the concept of a standard rather than progress as the goal.

Now, you and I know that statistically everyone cannot score at the median which in achievement tests most often is translated as grade level equivalents. Also, performance seems to vary from subtest to subtest.

I just do not know how -- even with all of us working together -- we are going to put across the fact that this child's progress is more important than a standard of achievement. Take for example, a sixth grade pupil who tests at the fifth grade level and who, in the eighth grade, tests at the seventh grade level. He actually is not up to the standard but he has made steady progress, which is lost sight of because of the imposition of a standard rather than progress as the measure of achievement.

Closely related to this point is the problem posed by the American confidence in numbers. When we try to explain that one has to bear in mind the standard of measurement, the idea is brushed aside, perhaps because the public has to believe in the certainty of a number and cannot tolerate the slightest slippage of confidence.

Other problems center around cultural differences in children, the result of deviations and language difficulties, but these have become issues and let me comment about these. In problems of cultural differences, the results of deviations, vocabulary and language difficulties first came to the fore with respect to our black population.

Everyone here is familiar with the efforts to develop culture-free, fair tests and the apparent import of altering results. It is a dilemma for us all that the response of the black community was and is to put pressure on the school system for a better job of teaching.

Now, the pressure is coming from the Mexican-American and Puerto Rican in the Chicago school area. Of these voices, the Puerto Rican dominates. They raise the issue not only of cultural impositions but also of expectations as to confidence in English, and they are firing at the testing program, which they say penalizes Puerto Rican children because of cultural differences and problems of English. We have found, of course, that many Puerto Rican children who cannot read English cannot read Spanish either. But I would add here that the minority communities do not want explanations as to why there is a lag or why we have not delivered.

They are saying I think rightly -- let us skip over all this and let us do the job now. There is really another reason that I am here. Several Board meetings back, one of our minority Board members pointed her finger across the room to me and said, "Mr. Byrd, what are you doing in the Chicago schools to see that the test makers are considering the kinds of youngsters we have to serve in developing the kind of instrument that will do a better job of measuring

their achievement and measuring their abilities? And I say to you, Mr. Byrd, whatever you have done has not been enough and if you haven't done anything you had better get started. You had better work quickly and you had better let other big cities know we are ready to combine with them to exert the kind of leverage that will result in getting the kind of instrument that we need to serve our youngsters."

This was the same Board member who had in mind that we had just recently taken a position, insofar as adopting textbooks, that no longer will we buy the best available but the books must meet certain standards. She was saying to me clearly that insofar as group tests were concerned, we had better have test makers hear us, and, insofar as individual tests were concerned, we are not concerned just with translating what we have got from the English language to other languages. We are saying that the test makers and those concerned with test-making had better take into consideration the various cultural differences of those involved and we had better set about the job if we are to continue.

What are the remedies? I touch very sketchily upon remedies, for the fire must be extinguished and the turmoil quieted. First, there is no question but that we need better tests -- better in content and format. They must be made more clearly relevant and more quickly responsive to current needs. I do not know how the last two can be achieved, but if they are not achieved I do not know how school systems like Chicago, for example, which is always on the verge of impoverishment, can in turn respond with the repeated purchase of new booklets, to strike a very practical note. The changing of tests frequently has other obvious disadvantages.

What I am saying is that one solution is to find a way to do the nearly impossible, but doing that merely takes a bit more time and probably a little more money.

Secondly, we need more sophisticated understanding, interpretation, and use of test results by teachers, counselors, and administrators. Somehow, working together, we all have to canonize those numbers and return to some confidence in our professional assessments and our professional judgment. In a sense, there has been an advocation in favor of numerals as the dictator, a handy fellow to whom to pass the buck of responsibility. Finally, all of us together have to put standardized testing back into context. For example, in 1968 there was published a handbook called "Guidance Service for Illinois Schools." This publication included a section headed "Guidelines for Developing the Testing Program," which states: "Keep in mind the test is but an indicator of a pupil's performance on a given day under a given set of circumstances."

Most of that quotation is printed in bold type. There are other remedies but I doubt if they are more in number than could be covered by these categories: better instruments, more sophisticated use, and wider prospectives of testing. I might add that in Chicago we are not only concerned, we are not only caught up in turmoil about testing of students, but in a big system such as ours, we are caught up in the turmoil of testing teacher applicants, of testing administrator applicants, and I have not lost confidence in the ability of test makers to respond to a need.

I call to mind an experience we are in right now. We talked with the test makers about developing an examination for the position of principal in our schools. We have just completed the written part of that examination. Some 700 candidates took that examination. Of the 167 who passed the written phase, some 43 percent of them were members of minority groups when only 40 percent of minority members took the examination. This is amazing as far as Chicago is concerned, for in this one examination we have had more black candidates pass the written examination for principal than all the examinations since 1946, and these examinations have been given every three years or less.

So there has been a response to a need and I think the publishers have responded. We are having this difficulty with teachers and paraprofessionals and they are saying -- and turmoil grows -- that the testing activities, text testing exercises, are not fair and must be revised.

I began these remarks by suggesting that stardardized testing as a part of the educational establishment had its turn in turmoil coming and, concluding, I return to that point. Just as the school is indirectly being held responsible for the results of social deficiencies in this country, it is also being trapped directly by a kind of overkill or oversell in testing.

I do not want to see a moratorium on testing, but I do want to see a better result. What I would like to see is a return to moderation on the one hand and to responsibility on the other. Test publishers have to moderate their oversell. Test users have to upgrade their insights and return to their responsibilities.

Test publishers have to assume more responsibility for ensuring adequate interpretation and appropriate use of test results with, for example, a better description of the norm group. Test users must moderate their reliance upon results.

In short, we have to douse the fire and quiet the turmoil by some united professional approach. Testing is lucrative, big business. Education is big business, but it is not financially lucrative. Education, is, however, not only the highway to the gross national product and dividends of separate companies, but also the gateway to the American ethic.

Neither testing nor education is isolated in turmoil today in this country or elsewhere. Newspapers, the radio, and television never let any of us forget that fact. Thus we cannot escape the crisis of the fire and turmoil. We can, however, work together to overcome our deficiencies and to bulwark our strength. Indeed we must. But we have to have the sense to discern the difference between them and the integrity to act on our collective discernment. What I have been trying to say these last few minutes is that testing is in the midst of a crisis, the new turmoil -- rightfully so. We as users have a responsibility, and the test makers have a responsibility, to resolve this crisis and I would submit to you that we don't have an eternity in which to resolve it; for, indeed, if we do not the voices that I hear, the pressures that I feel, say that either you do something about it or we will abolish it altogether. To me that would be catastrophic and it brings this to mind.

A story is told that a golfer went out on the green, teed up his ball, addressed it, and prepared to make a shot. He took a vicious swing, missed the ball, took out a pretty good swathe of turf in the process, and almost demolished an anthill. Unperturbed, he moved back and took another swing with the same results. The ball, unmolested, remained on the tee. Another swathe of turf and another big bash into that anthill. One of the ants, sizing up the situation, said to a surviving member, "You know, if we are going to get out of this alive we had better get on the ball." Thank you very much.

CHAIRMAN HUME: I would like to say Amen, Mr. Byrd, for this constructive presentation. I would also like to say that, in the 15 or 16 years during which I have been coming to these ERB luncheon meetings, this is the first time I have seen no one leave the room, either before or during the speaker's presentation. This gives you an idea of what we think.

I wonder if there are members of the audience who would like to address questions to Mr. Byrd. There are microphones located around the room. I think he would be willing to answer you if you have specific things you would like to ask.

MEMBER: Was I correct in understanding that your pupil population is about 500,000?

MR. BYRD: Yes. The number is 580,000. Of these, some 140,000 are high school students.

MEMBER: Okay. Was I also correct in understanding that you are giving four million tests per year?

MR. BYRD: How many? No, I certainly misread that. It is slightly over a quarter of a million per year.

MEMBER: Another question I had was why have you elected to use the median score rather than the mean? I as assuming you have a fairly large school.

MR. BYRD: I am at a loss as to answer, except to say that, when I looked at the median and mean, the median was better.

MEMBER: I have one last statement. You seem to be unhappy with the norm, as I think many of us are from time to time, but I am wondering if you do not have your own in-house equipment and computerization for scoring.

MR. BYRD: We are in the process of updating that kind of in-house service. We do not have it presently. I must say that over the last couple of years -- especially the last year -- we have had many invitations from test makers to participate in the norming process, and we are accepting.

MEMBER: Okay. But we have noticed that one thing that happens when you attempt to participate by becoming part of a norm sample is that, later, you are dropped. Most large districts, certainly those of your size, would probably have test equipment equal or superior to that of the test makers; therefore, in theory, you should soon be able to come up with a norm table of your own.

MR. BYRD: Well, to repeat what I said earlier. We are diagnosing more. We have made inroads recently and, hopefully, are continuing in this direction. Let me say this: I have in the audience our director of testing,

Dr. Elmer Casey, and -- back in the office -- when the real technical questions come up I call upon the person who is a specialist in that area.

However, as an administrator I have real responsibilities in trying to quiet the turmoil and put out the fires relative to anything that happens. So I am interested in everything that goes on, and testing is one of my big headaches.

MEMBER: You say that you use testing as the means of selecting staff. I would like to know why you do so, when so many communities are interested in mere interview and certification.

MR. BYRD: One of the reasons why we do it is that our attorneys tell us it is a requirement of the law. We have on the Illinois state statutes the requirement that communities of 500,000 or more must establish a board of commissioners to examine the teachers for fitness to teach. Therefore, the regular certification processes of the rest of the State of Illinois do not apply to the City of Chicago. Now, the Board of Examiners, for a number of years, has used this technique for establishing a list of eligible persons who are qualified to teach a measure of their performance on the test.

At one time it was an in-house instrument that was used generally. Now we are accepting results from the National Teachers Examination but we are using those results. Now, with the pressure building up, we have made some modifications in that, and our attorneys have reread the law and have found that a person, after successful experience of three years or more, may become certified through that route. I am only saying that as the pressures continue maybe there will be another reading of it and another modification. At this point, that is where we stand.

MEMBER: Do you have in your system (maybe the answer is one you would go to your director for, also) any sense that a part of the turmoil comes from expecting the same tests to accomplish too many different things? That is, after all, in your school business you are concerned with their use and administration in the schools, in guidance, with the work of instruction in the classroom and so on.

Now, to what extent can we set different prescriptions for the preparation of tests for different purposes? That is, is your Board working in this area?

MR. BYRD: I think our department of testing is working in this area. Certainly, we have a great responsibility in this area to revise our philosophy and use of tests and our whole approach to the subject. Thank you very much.

CHAIRMAN HUME: Thanks again. I now adjourn this session.

(Whereupon the meeting was recessed at 2:15 o'clock p.m.)

- - -

THURSDAY AFTERNOON SESSION

Session One

October 29, 1970

- - -

The "Building A School Testing Program" session of the Thirty-Fifth Annual Educational Conference of the Educational Records Bureau convened in the Grand Ballroom of the Hotel Roosevelt, New York, N.Y. Thursday afternoon, October 29, 1970, and was called to order at 2:30 p.m. by Chairman F. Martin Brown.

The Thursday afternoon session entitled "Building a School Testing Program" consisted of a panel which included Frank B. Womer, Daniel Wagner, Jean Garten, and Donald Roberts. Represented on the panel were an administrator, a counselor, a teacher, and a general measurement consultant. The approach was that of a consultant working with a school system to review the school's testing program, using a committee within the school to review, evaluate, and plan. The general goal was to develop a program that would give the school's faculty and administration more information about students and to provide students with more information to help them learn more about themselves.

The presentation was developed around the concept of a simulated committee meeting. The purpose within the committee was to define the objectives of the school's testing program to proceed with revision or development of the program, and to prepare a system for appropriate dissemination of test results. Each panel member developed a portion of the theme, and a general discussion followed reacting to the overall concept of building a school testing program.

Mr. Wagner opened the presentation by setting the stage for the committee session. The committee was to reevaluate the ongoing testing program in a hypothetical school situation. The school was defined as a prekindergarten through grade 12, coed, surburban day school having 500 students and a teacher-student ratio of one to ten. Furthermore, it was non-sectarian and non-parochial, located in an upper-middle class setting and generally college preparatory in nature. The constituency of the school was made up of actively dissenting students and parents. However, the curriculum was very traditional and school administrators recognized the need for a thorough, extensive self-examination. The purpose of this self-examination was to set up an environment in which the school would listen more to the students and less to the colleges, trying to be sensitive both to students and the needs of society, and to be innovative in meeting those needs, but not necessarily rash. Coordinated efforts were being made by the guidance counselor, the faculty, advisory councils, and administration to ask penetrating questions of themselves and of the system, in order to come up with answers that would provide new direction in working with the students. The faculty was depicted as well-trained, with a researching, experimental type of personnel.

The specific aim of this simulated committee was to look at the testing program in the upper school and review the objectives of the testing program. It was decided that this program should fit the philosophy of the school very closely. If parents and students were asking why the school presented programs as it did, then it was the responsibility of the faculty to be able to answer this question by giving a purposeful program as well as objectives. It was determined immediately that any measurement program should be used more qualitatively than quantitatively and test information and results used positively, not negatively.

The next question to be answered was, "Are we truly using our measurement program to individualize instruction, to motivate individuals, to give direction and purpose, and to help put students into society in the proper manner? Or are we using it in more restrictive and less enlightened ways?" It was agreed that a testing program should definitely emphasize individual learning differences. In the process of individual study, it was felt, more emphasis on a team approach between counselors, teachers, and administrators could be used better to meet the needs of students. The point was made that every student needs to have information about himself presented on a personal, one-to-one basis. Such an approach, it was said, can also avoid misconceptions produced due to labeling of students by teachers who do not interpret test results properly, especially by the use of test results in a way that does not correlate the information with other known behavioral facts about students.

Questions were raised in the simulated committee meetings about group testing procedures. It was stated that appraisals must be done on a continual basis to provide longitudinal information on student growth. Other questions discussed included the following. "Should we test everyone in the group or only the children with learning problems?" "Is our testing program for the identification of children with special needs, and are we capitalizing on this information by providing the necessary programs to bring about improvement in student achievement and skills?" The consensus was to work toward a general survey of achievement of the group, but to recommend appropriate diagnostic instruments for teachers wanting to do additional analysis of students where indications would show the existence of various kinds of weaknesses.

Mrs. Garten assumed the role of the counselor on the simulated committee and developed the following concepts. Any testing program should be an information service for students to get information about themselves. It should help in the establishment of realistic educational and vocational roles. And it should be a service to parents to help them understand that realistic goals must be set during the education of their children. Finally, such programs should be designed in a way that provides guidance to faculty members as they try to structure an optimal learning situation. Since group testing is a large part of the information-gathering process, it was noted, it is critical to remember that one must deal with individual data even though it <u>is</u> a group testing situation. Individual results must be interpreted wisely and supplied individually. The counseling office, Mrs. Garten said, is in a position to give definitive help to the administrative office and to the teaching staff, providing guidance to students as well as applicants to the school.

Mr. Roberts, taking the role of the teacher on the simulated committee, gave a teacher's view of how testing is regarded. He was quick to admit that teachers do misinterpret tests. But, he said, this is often due to the fact that teachers are not brought into the full discussion of test results and their use. Realizing the tremendous number of tests available and the kind of instruments that are used within the school system, Mr. Roberts expressed real concern about the ability of teachers to use all available tests. He also questioned the extent of their understanding of tests and wondered if their training was sufficient so that they could be of any

real help in evaluation. As a teacher, he was quick to point out that many teachers are suspicious of what tests will tell them and they are often not provided the guidance necessary to understand specifically what the tests do have to say.

Following the discussion by the committee members, Dr. Womer then pointed out that a test program is only a part of a total information-gathering system in the school. It always supplements other evidence. He questioned the impact of testing in the school and asked for comment. Replies from the simulated committee suggested that teachers suspect testing in the upper schools, and often resist using the information instruments may supply at that level. The counselor pointed out that the guidance office often fails to give teachers the kind of help they need by not providing needed descriptive information. Counselors could help by describing what a student is like, she said, how he got that way, and any genetic factors that may have contributed to a particular educational situation that he was in. Counselors should be the first to provide predictive information for teachers on the basis of information in hand. Counselors should be quick to be of help, she continued, and willing to evaluate themselves and the effectiveness of their services. Guidance is often too hurried to give good results. It is inexcusable for the guidance office to not assist the faculty and administration to use standardized test results as a vital piece of information in the evaluation structure, Mrs. Carten concluded.

Other concepts discussed dealt with evaluation of schools by regional associations and state departments of education. It was pointed out that standardized tests have consistently been used in this role, but that care must be taken to assure that the context in which such information is used is realistic and accurate. Pressures also have been created on the schools through external testing programs such as the Secondary School Admissions Testing Program, Preliminary Scholastic Aptitude Test, and the Scholastic Aptitude Test. It was pointed out that the prevalence of so many testing programs has caused a reaction from some independent schools whereby they no longer feel independent. It would seem more important, it was felt, to de-emphasize external testing and to emphasize an internal testing program that would help the greatest number of school children in making individual adjustments.

Assuming the selected testing instruments are well designed, panelists said, how well the instrument is being used to measure effectiveness must be asked. The emphasis in the educational setting should be on the testing of educational progress, not so much on psychological testing. It is necessary for psychological testing to be used in restricted circumstances, but this should not be a major part of the testing program. It was noted that testing instruments are not designed to measure ego growth, and schools must recognize the basic service to be rendered by a definitive testing program. In the process of developing educational systems, it appears important that the psychology of behavior should have greater emphasis and that teachers should have more exposure to measurement principles in their training program. We are relying too much on subjective judgment, panel members said, when there are certain basic objective types of measures available, even though they may have certain limits in their application.

The focus within the committee then turned to the development of a testing program. Based on the various objectives identified, the types of instruments were discussed, as well as the question of when and how to use them. And other questions were raised. Should local norms be provided? How should testing be organized for spring and fall measures?

In regard to the grade levels at which testing might be done, it was pointed out that there is probably little need in many school settings for standardized tests in grades 11 and 12. The comment was made that this group is already much tested and bored to death with it. Grade 9 would seem an appropriate time, it was felt, to develop an aptitude score for students who are evidently college bound, and then to develop a College Entrance Examination Board score prediction. This would give evidence to expected success on college level material and would give counselors early indications of student's academic potential. For students still having difficulty with the command of English and general communication grade 10 is definitely not too late to work with diagnostic reading evaluations so that students can be given the guidance and counseling that may assist them in overcoming reading deficiencies. In the developmental process as students go through school, it was pointed out that Grade 7 is also a good time for diagnosis of reading difficulties as students enter the junior high level. If is has not been done before, this may also be an appropriate time for aptitude testing, possibly using multi-factor types of instruments in the seventh or eighth grades, it was agreed.

To assure students the best possible guidance, the panel agreed that it was essential to examine reading progress as it had been measured through the early years. Reference to ability scores might be in order, but these would need careful interpretation. There would also need to be reference to achievement measures of basic skills, since this would furnish teachers the objective kinds of information that would assist them in developing appropriate insights into what they might expect from students. Because of concern for information at the junior and senior high levels, it was felt that the lower school testing program should be coordinated with the upper school. Testing results at all levels could then be studied longitudinally, with measurement information gathered systematically for teacher use, and fitted into a well-designed program. The teacher would still have wide latitude in selecting the particular tests most relevant to her teaching and curriculum, and to the objectives she has for her class. In-service programs that acquaint a teacher with the types of available instruments, and that also assist her in selecting the appropriate ones, would undoubtedly also be very helpful, it was agreed.

Much behavior in the academic setting is not interpretable by any available tests. The non-cognitive factors in a student's educational experience have never been given adequate consideration in research. The simulated committee felt there is a definite need for better understanding of the complex matter of social relations and how it affects learning and behavior. It was suggested that the school staff might develop experimental work in this field on a limited basis, since it is an area that is gaining more attention and research is now being encouraged.

The matter of how tests are used and how to involve those who use them came next in the discussion. It was agreed that before a test is given there is an absolute need for the teacher to have complete familiarity with it, to have some involvement in the selection of it, and to be aware of the specific measurement characteristics of the test. Students should be given an explanation of the test and why it is being administered. It was thought that open discussion with faculty and students about the needs and aims of a measurement program will clarify testing objectives. Panel members agreed this would have the effect of reducing redundant testing and some of the hostility often experienced. It was suggested that parents also be informed about the testing program, with the same needs and aims described in appropriate terms. Generally, it was conceded that a wide-open review of the measurement program with teachers, students, and parents could be beneficial if handled properly.

The opinion was expressed that some teachers feel testing interferes with learning and is, therefore, wasted time. This would indicate that discussion of the rationale for tests with teachers and students, a general sharing of information, can add value to the testing process and assure that only essential testing is done. The limitations of testing instruments should not be avoided in discussions with teachers and students. It can be clearly stated that a testing instrument represents only a limited measure of total behavior and also represents behavior only at a given time. We must guard against assuming that it gives information that is absolute or of permanent nature, panel members agreed.

Students should be involved in the testing process on the basis of the experiences that they must go through in the present-day educational environment, committee members brought out. Students are apprehensive that test results are used against them in college entrance procedures, rather than for them. They are also aware that testing often lacks relevance and because of it, show hostility to being tested. Unless there is good faith on the part of teachers and administrators in dealing openly and honestly with students, panel members felt, there will be a loss of faith in any objective testing program. More than ever, it calls for the school staff to be aware of its measurement objectives and to accurately define the minimum amount of measurement that needs to be done in order to get the valuable information needed for instruction and counseling.

The final phase of the committee discussion dealt with the dissemination of test results. It was agreed that testing must be repeated often enough to provide more than one measure of a particular type on a student, but not so often that test scores lose their meaning due to the proliferation of unneeded and unnecessary scores. The presentation of test results in language that is clearly understandable to students is an essential ingredient. The school testing program should be communicated carefully through well-prepared and well-written documents and by staff members who thoroughly understand the strengths, limitations, and weaknesses of testing and who know how to present them in discussions with individuals or groups of students.

In summary, it was apparent that the first task that faced the committee was the actual writing of objectives for a testing program. This would need to be done after a thorough evaluation of administrative, teacher, and student needs and an awareness of the nature of information needed. The second step would be the development of the testing program including the selection of test instruments, the time of year testing is to be done, the grade levels at which various tests will

be offered, and the particular students or groups of students to benefit from the program. The final step would be development of an appropriate and adequate system which provides for the dissemination and use of test results. This may include the determination to provide item analysis information for certain kinds of tests. It may mean that profile sheets with explanatory information must be prepared. It could also mean that there may be curriculum workshops in which test results are discussed relative to curriculum content and objectives. The teachers and counselors, it was felt, would have specific functions to perform in making the individual application of test results meaningful for the benefit of each student.

In conclusion, the panel felt the entire testing program must be a matter of continual sensitivity to children and the needs of society. It can be innovative, but must be practical and to the point. It can be extensive in its coverage of the entire school, but it must be limited in focus to actual needs. All in all, it was emphasized, a sound testing program can be a valuable asset to a school when it is administered by people who are aware and know what they are doing. When handled without insight and wisdom, it becomes a liability. The building of a school testing program requires an excellent discipline to assure that students obtain the individual attention that they deserve.

- - -

THURSDAY AFTERNOON SESSION

Session Two

October 29, 1970

- - -

The "Ethnic and Cultural Issues in Measurement" session of the 35th Annual Educational Conference of the Educational Records Bureau convened in the Oval Room of the Hotel Roosevelt, New York, N.Y., Thursday afternoon, October 29, 1970, and was called to order at 2:30 o'clock p.m. by Chairman Wellington V. Grimes.

CHAIRMAN WELLINGTON V. GRIMES: Good afternoon. It is my pleasure to welcome you to this session on "Ethnic and Cultural Issues in Measurement." But, before I introduce to you the Chairman for this program, I would like very much to read to you a very short letter which was written in 1744 when a commission from Maryland and Virginia was negotiating a treaty with the Indian nations at Lancaster in which the Indians were invited to send a number of their boys to William and Mary College.

The next day, after the invitation had been issued, the Indians declined the offer by letter as follows:

> "We know that you highly esteem the kind of learning taught in those colleges and that the maintenance of our young people while with you would be very expensive to you. We are convinced, therefore, that you mean to do us good by your proposal and we thank you heartily, but you who are wise must know that different nations have different conceptions of things and you will, therefore, not take it amiss if our ideas of this kind of education happen not to be the same as yours.
>
> "Several of our young people were formerly brought up at the colleges of the Northern Provinces. They are instructed in all your sciences but when they came back to us they were bad runners, ignorant of every means of living in the woods; neither fit for hunters, warriors nor counselors; totally good for nothing.
>
> "We are, however, not the less obligated by your kind offer though we decline accepting it. And to show our gratitude for it, if the gentlemen of Virginia will send us a dozen of their sons we will take care of their education, instruct them in all we know and make men of them."

I think maybe there is something here that we will have an opportunity to reflect on as the panel goes on this afternoon. Therefore, it is my pleasure to present to you Richard C. Kelsey, who will serve as Program Chairman. Mr. Kelsey is executive assistant for the Office of Non-White Concerns of the American Personnel and Guidance Association in Washington, D.C. Mr. Kelsey.

PROGRAM CHAIRMAN RICHARD C. KELSEY: Thank you, Mr. Grimes. I have gotten quite a few inquiries about the question mark in the program. To set the stage, I purposely did not send in information that could be printed here because today I wanted to present this whole session in a somewhat humanizing light as we begin to question some of the standards; and not only to question them and the credentials that go along with them but -- perhaps as a result of this session, or perhaps in the discussion -- we can go beyond that and begin to have some input that would suggest a plan of action.

I suppose for some time I have been listening to this whole business of controversy about testing but a few concrete suggestions from audiences and panels like this one as to what direction should be taken might help to provide solutions to some of the problems and issues about which we have been hearing all day.

Let me stop here and introduce the panelists. On my immediate right, here, is Paul Collins who is the director of testing at Washington Technical Institute in Washington, D. C. Mr. Collins will address his primary remarks to selection and placement as it relates to the ethnic and cultural issues.

And then Mrs. Joyce Hicks -- I might say that I want to emphasize Mrs. Joyce Hicks. I once fell in the trap of saying Mrs. Charles Hicks. In the whole business of looking at new types of definition, it is tremendously important for all of us to realize that everybody is caught up in redefinition, and women are beginning to express themselves and they must be considered, too, as a kind of cultural group we need to give attention to. Mrs. Joyce Hicks is with the Evaluations Section of the Board of Education of the Columbus Public Schools and will address herself to testing and its implications in the evaluations programs.

Next, Charles Hicks, who is a student. I thought it extremely important to have a student on the panel. He is in his Ph.D. program at Ohio State University and will address himself to the aspects of testing relating to cultural and ethnic issues.

Let me further set the stage and then say that, as I view this whole issue, it really is one that I might put in the framework of the oppressor versus the oppressed. You know, usually when you make that kind of statement people get excited about what you really mean.

What I really mean is that we have been boxed into certain kinds of standards that have been set by those who attempt to maintain the status quo, and we have been told that everyone must maintain these types of standards. I think we have heard that a number of times in different ways today. Yet, if we get boxed into that kind of interpretation, I think we tend to lose the significance of the individuals involved and we certainly tend to lose the contributions of the various subgroups.

We are apt to assign certain kinds of status symbols to individuals and groups on the basis of some phenomenon completely outside that group and, as particular members of this panel begin the discussion, I am sure they will hit on this a number of times.

I will now turn the meeting over to Mr. Collins, who will talk about selection and placement.

MR. PAUL COLLINS: Thanks, Rick. I don't know whether I should be pleased or not to be the first to make a presentation on this panel because I think that possibly the whole business of selection as it relates to testing should be reviewed.

Number one: human beings are in many instances more interested in hanging labels on other people than taking a good, hard look at themselves. That is number one. Why? Because people don't like to turn the lamp of introspection on themselves. What is really lacking in the whole business of testing is a method that is objective and is an accurate description of human and physical characteristics.

Phychologists, as we know, have tried for a long time to develop an instrument which would portray the normal personality. Unfortunately, the character of the individual, by its very nature, precludes the possibility of such development. However, it is more difficult, we find, to perceive this kind of personality model than it is to say what definitely constitutes a test of the intellect.

Even though a great deal of research has taken place, there is not yet available for general direction a test that will describe the personality of "normal people" with accuracy as found in the academic or the achievement, or the ability tests.

All major selection testing programs are designed to make students more competitive in the educational process. Until a few years ago, this was almost entirely relegated to just students. In the last two or three years we know that performance contracting has brought it into the classroom and has changed the role of the teacher as well as that of the student. The oldest of these programs, as far as testing is concerned, that we know about is one which is developed and operated by the College Board.

This program was started around the turn of the century. It was a result of a proposal that colleges which required examinations set a common examination which could lead to admission to a number of colleges for the students who took the examinations. The program has been used since that time

and its influence is now felt in most of the colleges across the country. But this changed as far as the College Board program was concerned as a result of a need; this need was for all schools requiring admission tests to set their own examinations. This resulted in the standardization of that particular procedure.

We know that around the turn of the century the schools which required examinations were schools with selective admission policies. We also know there were many other schools which did not require examinations for admissions. These were either private, small, church schools or, possibly black colleges which had a select clientele. They had no real examination requirements but the standard by which the College Board set up their tests became applicable to all those schools after the Second World War.

The 15 years following the Second World War saw a very great change in the American educational culture and this was a result of several forces: First, the G.I. Bill brought college education within the reach of thousands who could not have considered college without assistance. Not only was this precedent-setting, but also related to a change in the life style of a large segment of our population.

Second, the average income rose to new heights, making college education possible for children whose parents could not have afforded it a decade earlier.

Third, technology and business growth caused a new industrial revolution, thus stimulating the need for people with college education.

Fourth, college-educated people became more respected in the public eye and people wanted a college education because it was "the thing."

In addition to the program which was set up by the College Board, we do know that other testing programs developed and they focused on the student in transition. But, in addition, in the '50s the Westinghouse Talent Search spread across the country. Selective service examinations provided draft deferment for able students and students who were interested in receiving scholarships took all kinds of examinations for private industry, philanthropic scholarships, and so forth. The National Merit Scholarship, which is the largest one, also reached into a large number of schools which before had very little interest in it.

Now, selection by testing. I brought in that little bit of history to show that selection by testing is not new. It is not new at all. It has been going on a long time, but a great many people have been unfairly discriminated against by these tests. This is not only true in education, which starts long before the child reaches school, but it goes straight through from there into his job life and onto the career ladder. Every person here is evaluated daily by some sort of test, whether it is his employer's evaluation or whether it is a test he takes for a promotion. Whatever the case, people are always being evaluated by tests.

I proclaim that individuals who, because of certain environmental conditions -- whether they are ethnic conditions, or whether they are racial conditions -- who have been subjected to the type of standardized tests that we have given in the past decade have been unfairly discriminated against.

I do not say, as did the speaker at lunch, that we should declare a moratorium on tests. I say that we should provide for adequate facilities to upgrade the standards we use. Second, we should make sure that these are used for the purposes for which they were intended. Third, we should develop tests which evaluate and are valid for ethnic and minority groups. Fourth, we should utilize cultural measures which will not unfairly discriminate against those people.

PROGRAM CHAIRMAN DELSEY: After the three presentations, we want to have about ten minutes for interaction with the panelists. There may be interactions occurring here and then we will throw it open for the audience.

MRS. JOYCE HICKS: The process of evaluation, as we know it, has several aspects: 1) the attempt to assess a particular educational project; 2) an effort to see whether it is necessary to recycle, if there are some alterations that might be made in order to enhance the projected program; 3) an effort to see whether it is necessary to demolish the entire program or effort.

So, since the great emphasis is on the evaluation of educational programs within public schools, much of the evaluators' time is spent in assessing student achievement in a particular subject area or in the entire curriculum; or in an effort to assist decision makers in deciding just what to do with particular projects -- remedial projects, and so forth.

An evaluation of teaching methods that might have been used to go along with new and innovative projects is also made, or it might be just the comparison of one student's achievement in a particular geographical location with that of other students in this particular location. However, this is usually done through the use of some kind of standardized measurement and, in many instances, the results of these, when interpreted to the decision makers, have led them to raise questions such as: 1) Do minorities have the mental capacity actually to participate in the educational process; or 2) Do they have an aptitude for school learning. This is not necessarily because the students do not possess the aptitude or ability to achieve, but because of the interpretation of test results which, at this point, are not necessarily geared to the experiences of the minority student.

Most researchers do not accept the doctrine of innate mental ability or differences between ethnic groups and they usually try to relate any differences shown back to the environmental differences which play a considerable part in how a child achieves in school or how a child actually relates to the academic environment. Consequently, many educators and parents are attempting to challenge what we might call the usefulness of tests and measures for cultural minorities. Granted, there has been some experimentation with new types of tests and with different criteria for the determination of equalities, but better measurement instruments should be instituted in order to assess which might be called educational potential and also to test the performance of a particular student in any academic setting.

We find, in several evaluations, that the greatest need is for some kind of instrument that not only measures what the child actually does in the school settings, but what kind of influences are going on from the outside that allow this child to achieve or not to achieve at a certain level.

Data from a large number of studies comparing performance of culturally different students with those of the predominant culture on standardized tests of intelligence, achievement, aptitude, et cetera, demonstrate different ways in which a student finds himself substantially low in certain areas and relatively high in others.

The widely discussed Coleman Report documents still further the extent of disparity in scores of children from minority groups with those of other children on a variety of achievement and ability measures such as: verbal ability, nonverbal ability, reading comprehension, mathematical achievement, and general information in the natural sciences, social sciences, and the humanities.

One fact to consider in evaluating is that, even though the standardized test, or whatever measuring instrument is being used, plays a considerable part in what may be concluded about a particular student or particular subject area, we find it difficult to blame testing instruments in total. Several studies have suggested that the educational system within itself also fails students of minority groups and it is reasonably stated that by the twelfth grade, after a student has taken a series of tests, he is approximately one standard deviation below that of a student of a predominant group. In many instances, this relates back to the school setting because, in a test of mental and motor skills of infants it is shown that differences between the minority and predominant groups at the period between birth and 15 months have the same or similar mental abilities with slightly superior motor skill ability among the minority children. However, during the first year or at the start of school, there is a decrease in the minority child's achievement level and the relative disadvantage of minority group children seems to increase over a certain period of time as a consequence of the differential in school, family, cultural, and environmental milieu -- accompanying poverty, slums, racism, and other influences. These forces seem compelling in view of what we know about the effects of social and environmental factors on intellectual growth.

In view of this, an alternate hypothesis, I would say, would be that standardized tests developed to test ability and achievement are, in many instances, biased or, more reasonably, that present tests are so constituted that a very substantial portion of differences between minority students and those of the majority culture is associated with factors unique to those students of the majority culture.

What I am trying to say is that most tests are actually geared to the experiences of the predominant culture, and experiences which are unique to those students are not necessarily familiar to students of a minority culture. Consequently, minority students score poorly on many tests.

The subject of test bias is too complex to go into at this point; however, the question has been under investigation for quite some time and it should be noted, for clarification, that it is not my intent to suggest that the difference between the way two groups of students score on particular tests determines whether the test is biased or not. It merely suggests that the hypothesis should be examined.

As long as we are concerned with the problem of what might

be considered nature versus nurture or insofar as tests come to acquire what might be called a "societal function," it seems inevitable that researchers and test developers should seek ways of removing culturally linked variance from tests.

Some researchers have identified three approaches to this. One is compensation. This is a procedure in which the items known to favor one group or another are balanced so that the means of the group are equal.

The second is elimination. That is a procedure in which one eliminates items or types of items for which differences between groups occur, thus moving toward a "culturally free" test similar to the Davis-Eells test.

Another is the identification of new intellectual factors -- or new ways of measuring those factors -- offering promise as assessment of important psychological functions but without the significant evidence of socio-economic bias. This can be seen in the various works on elementary learning tasks.

The effect of culture upon the student's performance, particularly on intelligence tests, has been an area of considerable activity among psychologists, sociologists, and researchers. Perhaps no other factor has attracted similar attention, and rightfully so, in the nature of testing. The responsibility of finding a so-called "culturally free" test has also received attention from numerous individuals concerned with education who have an interest in the testing field.

F. L. Goodenough best summarizes this in her article in the Psychological Bulletin. She expresses the opinion that the search for a "culturally free" test, whether of intelligence or artistic ability, personal or social characteristics, or any other measurable trait, is an illusion; and that a naive assumption that mere freedom from verbal requirements renders a test equally suitable for all groups is no longer a valid assumption.

In conclusion, as a researcher I tend to view tests as cultural artifacts and feel that failing to take this into consideration could lead me or any other researcher to some erroneous assumptions, because if a test can predetermine a student's ability to succeed in school, this further reinforces the notion of interdependency of a test upon culture. And this is especially true when we consider the school as a cultural institution. Because of cultural differences it becomes almost a hopeless task to attempt to measure differences between ethnic groups with presently available tests.

Kleinberg states that the variety of attitudes and points of view which we collectively call culture may produce such different reactions as to make direct comparison of two cultural groups scientifically insignificant. Therefore, the role of the cultural factor and its effects on measurement instruments cannot be overemphasized.

When examining other influences on test performance, consideration must be given to the following: the language patterns of the particular student in a particular culture; what motivates this particular student to achieve or not to achieve in school; the rapport the student has with people within the academic setting; and the physical factors of the student (sight, hearing, etc.). The race of the examiner could also have considerable influence on how students perform in a test. And cultural influences of the home and the neighborhood are other considerations. In trying to find certain tests, we must be aware of the fact that present criticism has somehow obscured the real problem of measurement and what we must do is refrain from becoming too critical of a particular instrument, but instead, find one that will achieve (at least partially), what we are attempting. Therefore, tests should be considered as generally useful for a limited number of strictly practical purposes, and major improvement of these instruments should be given priority by test developers and researchers. For the present time, tests should be looked at critically with notations being made on every report of test results that the particular test only measures a given aspect. Thus, we eliminate the assumption that the test is a complete measurement on which all facts relative to decision-making in this area can be based.

Thank you.

PROGRAM CHAIRMAN KELSEY: Thank you, Joyce.

MR. CHARLES J. HICKS, JR.: My wife and I had originally planned that, if needed, I would share some of my time with her.

In conclusion, or in addition to what has previously transpired, I would agree that tests and some reliance on tests are a fact of life within our society, its institutions and agencies. We are indeed a quantitative society whereby crucial and vital decisions are made, based on such factors as: what the score was; who had the highest score; what the qualifying scores are; what the cutoff levels are; and so forth.

A person's performance in a test reflects the degree of success of his acquired training and the unsuccessful and faulty attempts by the educational system. Therefore, tests don't actually, as we say, test people as much as they indicate the failure of a system to cover all of its members -- including those of minority groups.

Differences in test performances by people from different ethnic, social, and economic levels are due to the diversity of cultural experiences. These differences are reflected in terms of behaviors, attitudes, expressions, experiences and habits displayed in nonconformity with the particular social cultural expectations of the predominant society.

The differences that exist between the various ethnic, cultural and sub-cultural groups reflect the divergent socialization processes determined by forces within our society. Standardized tests fail to take these aspects into account. Consequently, members of minority groups are victims of quantitative decisions arrived at via testing instruments and generalization of results. On these bases alone, decisions are made which perpetuate the negativism experienced by these groups throughout their daily lives.

"Standardized tests," applied to minority groups, depict the differences between these groups and the majority; and those differences are viewed by the predominant society in a negative light. Thus, "these people" -- as they are sometimes called -- are proclaimed culprits and, in being so proclaimed, they are condemned, isolated, alienated, and debilitated.

To the average person from a minority group, a test is just another experience to be viewed as punishment in which his "weaknesses" are shown; and his unfitness and illegitimacy are defined. In essence, it is an experience which points out weaknesses, shortcomings, inadequacies, insecurities, and failures, as indicated by the misuse of the test.

My thesis is that -- within the context of our times, and taking into consideration the influence of historical-social-cultural forces -- tests, testing, evaluation, and appraisal efforts tend to lead to the debasement of man -- particularly members of ethnic and cultural groups that differ from the majority or the predominant culture.

The fact I want to impress upon you is that our "Great American Society," its culture, and all of its methods, techniques, tools, institutions, and so forth, are at this point in time debilitating man's capability for realization of the essence of being "human." And I use the example of the minority groups only as an indication of that fact.

These tests do, in fact, contribute to the debasement of man -- and by debasement I mean the lowering in status, esteem, and quality of character; the reduction in position, worth, value, and dignity; the destruction of purity, validity, and effectiveness; the definition as illegitimate or deficient; the causing of moral deterioration, twisting and distortion, depression, degradation, and injury to social standing; the wounding of a person's pride, causing deep shame; destroying self-possession and self-confidence; the beating down, nullifying, and reduction in degree or intensity of the value of existence itself.

PROGRAM CHAIRMAN KELSEY: You know, it seems to me I have heard a number of various points of view as far as what is happening. I think it is tremendously important to identify that has happened and what is still happening. Let me summarize what I think I understood the three speakers to say.

First: I think they were saying that testing, as we have typically been using it, is what I like to call "administrative expediency," and almost eliminates the whole business of what we like to say about the humanizing of individuals. We operate basically for the continuation of a particular institution, to make sure things run smoothly; thus we develop a kind of norm and we force ourselves into a cultural straightjacket, reacting to one kind of culture only -- if I interpret correctly. That is one of the factors I think I heard discussed. We talk about multiple cultures within the American society; yet, we do not respond to them within the educational system.

Second: I think I heard Joyce talk about the necessity for building cultural links, and I am not sure we have addressed ourselves to this. I hope, in the discussion, you will address yourself to some possible solutions to this factor in connection with testing. Concurrently, with that, let me say I have a feeling we are all aware of some of the ills. We are all aware of some of the misuses, but we have never really dealt with them or planned a course of action. We tend to say such things as, "Yes, I recognize weaknesses but those are the best methods we have for measuring," rather than

asking ourselves what we can do to enhance each individual, which is what we claim we are all about.

Let me throw the discussion open again for reaction back and forth between the panelists and then I will open it up to the audience.

MR. COLLINS: I heard Joyce say that in the administration or evaluation of tests, the race of the examiner could have some -- what did you say?

MRS. HICKS: That race could have an effect on the child's performance in the test.

MR. COLLINS: Would it not be a question of the conditions under which the tests are administered? Whether a child has a proper frame of mind, whether the person is objective or racist? Why would it necessarily have a bearing on the performance of the individual? Is this based on research? I think I have heard this statement before.

MRS. HICKS: Well, yes. From what I have gathered from two or three studies (one was, I think, in the Journal of Negro Education, Volume 37, 1968) I remember seeing an article describing research on the race of a particular person not only in testing, but in classroom situations or whenever a student has to interreact with "an authoritative" figure. The study in JNE reported that race does have some influence on the way the student performs, not necessarily in tests but also in classroom activities.

Now what do you mean about the right frame of mind? I do not quite understand if you mean that a person has no race bias -- then I can not really respond to that because I wonder who is actual y free of racial bias.

MR. COLLINS: --- and whether we are speaking of overt racism?

MRS. HICKS: Not necessarily. Personally -- I hate to give personal references -- I respond differently to a black than I do to a white if I am in a classroom setting, and it isn't totally against this particular person, but I recognize the fact that this person is here and I feel more togetherness or connectedness with another black person, whether I know him or not, until he proves otherwise. And that is only because of the division within society that has pushed me to this point where I am actually identifying people according to race until I learn differently. Once I get to establish rapport, maybe it won't make any difference at all. But how many students get the opportunity to establish the kind of rapport I am talking about with the examiners?

MR. COLLINS: I agree. I agree that in any situation where you test a large number of students in the conditions under which tests are administered in the public schools, it is a very amazing thing that students score as well as they do. Usually conditions are poor. The students are not motivated. The examiner probably is a person who does not have one bit of faith or confidence in the test; and doesn't know why he is giving the test but feels that he could be doing something better with his time.

Students are not prepared and are usually herded together in some cafeteria that has all kinds of bad things going for them. People are walking around, looking over their shoulders, and this is especially true in the inner-city schools. I think we are all aware that the inner-city schools are mostly populated with the disadvantaged, whether culturally, economically, or racially. These are the persons usually herded together with the attitude that this is just another chore. We are not going to do anything. We are not going to change anything; most of the time we are not going to use examination results; but, whenever examinations results are used, they are generally used to the degradation and detriment of the people examined.

If anything comes out of this discussion it has to be the fact that tests need to be made more relevant. When we talk about relevancy, we are not necessarily saying that all of the items have to be changed, but we are saying that persons have to be made more aware of the possible good uses that can come from test results. We mean that there must be a better interpretation of the test results, and they must be used. If they are not going to be used, then I say -- as someone else has said -- declare a moratorium on testing.

MR. HICKS: I think tests are very relevant to the times. I think they exemplify the kind of debasing procedures that are constantly evolving in our society. To that extent they are relevant.

PROGRAM CHARIMAN KELSEY: You presented another issue, too, that has to do with the race of the examiner. Among many members of sub-cultures within the economy, I would suspect that -- in looking at a white person -- there are, basically, two types of model. Either a person is overtly a racist, or what some people might call a "liberal racist;" thus, in the minds of the students he examines, he is likely to see some kind of expression. I think that is what Joyce was getting at. This is as critical a consideration as the test items themselves in respect of the value, or lack of value, of tests.

MR. HICKS: May I say one thing? I am not interested in who administered the test. It is just that this was pointed out as one thing that might hamper a student's ability to respond.

MEMBER: Mr. Collins brought up something to which Mrs. Hicks responded. The point is that the whole business of racism has been brought up, and my question is whether or not we can translate this into some kind of process that we can deal with in a more objective and rational fashion.

PROGRAM CHAIRMAN KELSEY: This is one of the things I am extremely concerned about. We have typically tried to respond to issues in a somewhat neutral fashion, rather than dealing with the real psycho-social involvement; thus we have not come up with a solution. I would suggest that one way to get at the solution is to deal with realities, even though this may hurt. We used realities as a medium through which to talk with you today. The test is really not the basic issue. There are other things we have not dealt with.

MEMBER: I have two questions. I want to hear more about the cultural link you mentioned and, also, I wonder if there is any doubt that tests could be created that would be fair to cultural minorities. The next step is what to do after we have them in school? How do we prepare them to take their place in the majority society, to have a fair place in it? Sure, we can make tests, but then, after we get fair tests, how do we get a fair education?

MR. COLLINS: First, I think the thing we have to realize is that we are not saying: do not give tests. What we are really saying is: if you are going to give tests, then do as you have been doing all along. As you know, the business of competition has been one-sided. In many instances it was not required for this group to take tests in order to go somewhere -- because they were not going anywhere.

Second, we used these tests, and not only did we use the tests to get people admitted into college or private secondary schools, but the teachers geared the curriculum to preparing people to take tests, because everybody knows that a person just doesn't chalk up a score of 600 - 800 on College Board unless they are prepared, and we do not mean that they are taught how to pass a test. We are talking about people who have taken the tests, who have copies of the tests, and who gear the curriculum to them -- oh, yes, and people who teach around this kind of test. Now, we have just said that, in persons from the age of birth to 15 months, there is relative equality and potential, but from that point on it is a matter of environment. It is a matter of nurture, heredity, whatever it is, developing in that person the potential to do whatever God created him to do.

In some households education is not a byword. In some households magazines are not even available. In some households people try to raise kids on $155 a month welfare. You know, this is what it is all about, and we have to address ourselves to these other issues.

You know what it takes to get into a "seven-sister's" school. You know what it takes to get into one of the "big ten schools." And you know that people in English classes -- 11th and 12th grades -- prepare kids to take the test. It is just getting to the point where we must either do the same thing for all kids or give it up as a bad job. Some of the larger schools have recognized this fact and have initiated projects whereby experimental groups come in without the benefit of College Boards and the schools give remedial training and hope the students will take their place in the larger society and be able to make it. These kids do make it, despite handicaps. They make it because they have something in them called "pride in themselves," because somebody told them a long time ago, "no one is any better than you." They make it despite these handicaps. They will eventually get where they are going, but this is in spite of the system, not because of it, and what we are saying here is that, if there is to be an end of turmoil in testing we shall have to give the whole society the same kind of human treatment.

MR. HICKS: Subhuman!

MR. COLLINS: Okay, if that is what it is, because if you prepare a kid for a test, he is going to say, "I have to make this high score..." This is the same thing that happened on Wall Street. It happens in every big business promotion. This is not the cause of the system; this is a result of the system.

MRS. HICKS: In response to your question about cultural links, check Horrocks' Assessment of Behavior, educational research journals, and also Educational Index under "Testing Usage."

MEMBER: In keeping with what Mr. Collins said, I would

like to suggest that it is patently unfair to keep a group of kids in school from 9:00 to 3:00 and then tell them they are labeled disadvantaged. They stay after school in order to get help from the same people who are not helping them from 9:00 to 3:00.

I think, if we are going to use Title I money to advantage in determining testing procedures, we should make the directors accountable. We should develop experimental methods and crank them into the 9:00 to 3:00 program.

Why keep kids who are disadvantaged, who are different, in limbo from 9:00 to 3:00 as discipline problems and then keep them after school? The fact is that most of the kids in the Title I programs are not the ones who should be involved in the first place because, if Title I programs are voluntary, the kids who are really the hard core are not going to elect to give up baseball or anything else to come in and listen to the same teachers.

MEMBER: I would like to make a comment and ask a question at the same time. I seem to get two different trends flowing out of the panel.

PROGRAM CHAIRMAN KELSEY: That is intentional.

MEMBER: One trend seems to be concerned with the process of test-taking in terms of instruments, and they are to make "standardized procedures" more standardized, so that available norms become more interpretable in terms of all students taking the examination. The other trend I get suggests that the tests in themselves are bad.

In dealing with the second opinion, I would like to differentiate between two types of tests: those that are concerned specifically with course content as it exists in the school (and there are some fairly specific criterion references here), and those that are basically concerned with evaluating a test, whatever criterion the test utilizes. If you don't accept that, you can't accept the test. Suppose we are interested in teaching and measuring mathematics, for instance. Shall we say there should be different criteria for what represents successful performance in mathematics? You cannot accept a test if you do not accept the criterion.

If you go one step further and look for culturally free instruments (which do not exist, except as a generally accepted approximation) you find that the closer you get to that objective the less those instruments reflect what happens in school. This is because of what you are pulling out of the test when you go to these. You are pulling out part of the test overlapped with a criterion. That may be fine if you say you do not approve of criteria. If you accept a criterion representing today's mathematics in schools as being what it should be, then you want a test that will reflect this.

What are you going to do with a test that doesn't reflect this criterion? The point is, if you accept criteria, then you have to be concerned with the processes for incorporating greater fairness in the test-taking procedure and the application of the norms to all those who take the test.

If you are disturbed by criteria, then you are questioning the criteria that underlie the subject areas being taught in the schools today. I get these two different currents here: one seems to be related to what kind of people we are talking about, and how we can remediate or bring them more in line with the criteria that we largely deem useful in terms of the functions of our society? The other seems to imply that we are not concerned with conditions as they exist; we are concerned with differences -- and that is why we need different instruction!

PROGRAM CHAIRMAN KELSEY: I will respond to the first trend because, as I hear you talking about remediation, it automatically means something built in that is wrong with the person.

MEMBER: No, plenty of people supposedly go through an average experience, where everything looks all right, and would be somewhere in the middle of the class. Something goes wrong along the way where they should be performing better, should be able to get more of the advantages of the school situation but are not doing so. We try to look for possible weaknesses and try to find ways of remediating. You do not seem to like that word, but the word is applicable for all groups. Anyone who is having problems needs remediation of some sort. Very advanced students who are well above the rest of the class need remediation in order to perform at the level where they should be.

PROGRAM CHAIRMAN KELSEY: I understand very well about remediation, but the point is that, as we use the word "remediation" I have the notion that we are talking about a set of values to which all must subscribe. You cannot evolve another set of values because if you excel in this second set of values and don't measure up on the first set, then you need help, and thus I say we need to reexamine our criteria and maybe even reexamine the whole idea of remediation.

MEMBER: Within a certain area certain things are basic. If you are going to build a bridge you have to know elementary mathematics.

PROGRAM CHAIRMAN KELSEY: That is a pure assumption. Under the present value system we assume that.

MEMBER: If you don't make certain assumptions you will have, as Dr. Appley says, to go back and rediscover the wheel; otherwise, you will have everyone wondering whether -- if all things are equal -- then, there is no standard in that world, no way to measure the quality of things.

PROGRAM CHAIRMAN KELSEY: I guess I am just resisting one standard, not all standards.

MEMBER: Oh, no. The basic criteria of performing deal with reading, writing, and arithmetic, and those are the main concerns of measurement in schools.

PROGRAM CHAIRMAN KELSEY: I think someone else wants to respond to you back there.

MEMBER: I am a little confused, too, because I teach in a predominantly black college where youngsters want to be accountants but cannot add two and two. We do not consider that debasing the student at all. The student knows that, if he wants to be an accountant, then he has to achieve a certain level of elementary arithmetic. We have to do remedial work. Can you tell us what else we could do? He insists he wants to be an accountant.

MR. HICKS: I was going to say that, if you are offering him remediation, then he has already been debased. The process has already been successful.

MEMBER: I disagree with you. He has come to us because he wants to be helped. He couldn't get into other colleges because of the kind of education he has had ...

MR. HICKS: He has been debased already.

MEMBER: I do not agree with you. If he were debased, he would not be coming to us to try to get an education.

MEMBER: Take music, I cannot sing a note but I don't feel debased!

MRS. HICKS: One minute please. This is totally different. It bothers me -- I am sorry. I got totally off the whole thing.

MEMBER: I was speaking of my college -- I could name the college because we feel we are doing a job, helping people.

MR. HICKS: Your particular college? I am sorry.

MEMBER .. at which I teach. I did not make any general statement.

MEMBER: What I would like to do is to try to pull together a couple of things I think you said, to fit in with some of the things I believe in, relating to this particular area.

First of all, I think the major point is one we cannot miss. If we do have a basically racist society -- which I think can be demonstrated -- any process we develop to perpetuate that society picks up a lot of institutionalized elements of racism.

I think, if we look at testing as one part of this whole, white, racist society, we do have that kind of process. I think, if we then look at the criteria which have been brought up recently in simple terms of standards reflecting what goes on in our society, if we do have a white, racist society, we will then develop tests that, in turn, reflect those particular criteria, so I think most of our testing procedures pick up attitudes of race along the way. This is why we have some of the problems that we mentioned, where many black people simply regard the whole process of testing, regardless of content -- whether aptitude or other type of test -- regard the whole process of testing as something that is used against them, something that is negative. Therefore, we may have to take a good look at the whole process of what we call testing, and consider whether it must now be overhauled or possibly abandoned because we have developed so many channels of racism along with it.

I think those of you who have some idea of the differences in black caucuses, going on now in this country, should allow blacks who are interested to participate in the white society as accountants or in other capacities, to express themselves on those criteria and take those jobs, if they wish to do so.

However, I believe that if many black people simply do not want to continue in a society that it is going through testing and all that goes with it, we as white people in the society should try to create a situation where a black person is not simply obliged to join with the white society and reflect all of the criteria if he does not wish to do so. That summarizes my position, more or less.

MR. COLLINS: What would you suggest, sir?

MEMBER: As a solution? My suggestion (and I think this is something that, basically, black people will have to do) is to try to do something in my area and following their own leadership because, as a white person, there is not much I can do about this whole thing. What I mean to say is that I am not black and I do not know the culture and I cannot run down to a black neighborhood and develop a standard for black people. I think, in this particular area, for the development of unique procedures and criteria, for those things demonstrated as being useful to black people, they will develop their own instruments and their own society.

PROGRAM CHAIRMAN KELSEY: This lady here has been trying to say something.

MEMBER: I think we have confused a number of evaluations with valuations; and valuations (I would say), are associated with the curriculum and with political processes. I wonder whether this gathering, concerned as it is with evaluation (which is an instrument of a particular culture with which I think a lot of us are dissatisfied) transcending both black and white people, can survey curricula in an effort to devise a relevent curriculum for all people. I do not think we should confuse schooling with education; or that educational potential should be equated with human potential, and I think there has been a considerable amount of confusion between the two. It is this confusion that has brought us into this ridiculous position where we talk about evaluation when the subject should be valuation. What do we value in society?

MRS. HICKS: I attempted to talk about the testing instrument in relation to educational evaluation, which has been defined as the assessment of a particular educational curriculum or subject matter, et cetera, in an effort to enable those who are responsible to make better decisions on particular educational issues. Testing was introduced simply because we utilize testing instruments in order to get at some of the points we are looking for, but I am talking about such matters as context -- input, output, obtaining information on a particular group, and that is evaluation.

PROGRAM CHAIRMAN KELSEY: Somebody else wanted to respond.

MEMBER: We should consider more seriously the importance of outcome in our society and the question of an outcome that has traditionally been important in schools -- to a degree that is out of proper proportion to its importance in society.

For example, even if a person is not going to be an accountant, it is altogether important that he should be able to read, to read charts and tables and graphs of the kind that you see in the newspapers. On the other hand, take the matter of English usage; as a result of the background in which I was brought up, I still cringe somewhat when a person uses a double negative. On the other hand, every foreign language I have been taught requires me to use double negatives as the correct way of expressing negation in the other language. There are things that are really important, that everybody needs in order to get ahead.

MEMBER: I have a number of points. One, I think there is misconception about the origin of testing. Tests were originally developed in Greece. We did not suddenly develop tests in the United States.

The Binet test was developed in Paris. Okay, let us start from there. There are approximately 210 million people in this country. One of the gentlemen seems to be advocating the qualitative kind of assessment. Let us assume that about ten percent of 210 million people are children. That is 21 million people.

Do you want to use a qualitative type of assessment? It just doesn't seem feasible. At least, the quantitative kind will give people all over the country a common nomenclature by which to judge and compare people. If I were an admissions officer at Columbia and you sent me 300 applications with qualitative judgment, how would you make an assessment, really? I would like someone to answer that point first.

MR. HICKS: I couldn't grasp it. Refine it a bit more.

MEMBER: With the number of people we have in this country you cannot get away from quantitative assessment.

MR. HICKS: Okay.

MEMBER: I would like to answer that. You were arguing for qualitative. I am saying that is not plausible.

MR. HICKS: I would agree that at this point in time in our society, there is not much quality in being a human being. I guess quantitative measures lead us in this direction. I would agree with you.

MEMBER: And another point -- I would like to finish my argument please. In any job there are certain numbers of requisite skills you acquire by doing job analysis.

MR. HICKS: I would agree with you that, in our society, certain things proclaim you somebody. If you have not taken on those skills and attitudes you are nobody.

MEMBER: I am not talking about attitude, per se. To become a physicist....

MR. HICKS: A physicist is not a "name" to somebody. It is a person, not just a thing -- or is it?

MEMBER: A physicist is a person who has demonstrated certain information in a number of germane areas in his field.

MR. HICKS: How do you differentiate a physicist from a person?

MEMBER: A physicist is a person with a number of skills. Would you want to be operated on by someone who did not have the requisite ability? Even if he were the most humane person in the world but did not have the requisite ability, would you want him to operate on you?

MEMBER: Last year there were 2.5 percent of black students in American medical colleges. Seventy percent of them have gone to Howard University for the past 60 years, according to research done by the president of Heary College in Tennessee. All of the rejects on the American Medical Association aptitude tests were accepted by Howard because these students do not get into Harvard, Yale, Princeton, or Stanford.

Over the past 75 years, these black medical students who couldn't make the grade completed their medical training successfully, passed through the state board and if you read the medical literature over the past 30 or 35 years, you will find that black researchers have contributed to important advances in medicine.

Black science students at City College this past year doubled in number over students in white medical schools. It is now 4.70, I believe. Most of those students do not have the qualifications required by the American Medical Association aptitude test. I dare say that, in the same proportion, white students who have qualified will also complete their courses successfully.

MEMBER: This is just my point. As has been demonstrated very clearly by the number of skilled people in our society. On that very basis there is some efficacy.

MR. COLLINS: You do not mean I.Q. tests?

MEMBER: There is a wealth of information stating that current I.Q. tests do not correlate very well with anything and this you will find in any psychological journal. In one of the most significant pieces of research being done right now in the United States, successful doctors indicate that the C students in college on all the criteria did as well as the A student in the world, and I.Q. helps predict school success but doesn't say a damn thing about what you can do outside school.

MEMBER: I entered this discussion with a great many prejudices that are not particularly the ones you might expect because I am a white person. Let me suggest that, in the matter of evaluation, it might be profitable to take a look at the September Ebony Magazine, in which the whole issue is given over to a discussion of some values between separation and integration, and I think there was something in the nature of ---

PROGRAM CHAIRMAN KELSEY: --- liberalization.

MEMBER: --- the point being that we are living in a world where we have to consider the facts of life as well as theories as to what we would like it to be. And these exist in the presence of the same set of facts that support all three or more of the views as to what we ought to do about it.

I have another prejudice. I think, if we are going to arrive at a little less turmoil within which testing becomes insignificant, there must be some equality. Some people are more equal than others! But there has to be some equality in the participation of the minority, ethnic groups, races,

and whites who happen to be the dominant group according to the facts of life -- and that means we cannot be shoving off all the problems onto the minorities. Instead, we have to share the problems with them. There needs to be communication among the groups that are in conflict. Communication is the start.

Sooner or later we may arrive at something approaching a consensus. Another thing we need is a multiplicity of acceptable outcomes, so that we do not have to have only one integrated set of values that will govern the whole damn society. We must have sets of values which fit the people who adhere to them. (Lord, this is almost free religion.) And these must be acceptable to the ones who do not adhere to them. Ours is a big society, in which we have many component groups whose values are compatible, but not identical.

Another point about this testing business: relevance, relevance to what? Now, I think it has been correctly stated that we are concerned with evaluations that help us to make decisions. And these decisions can vary all the way from what we see in youngsters now, to what we want them to become, or what we try to have them shape their behavior to within the next week, the next month, the next year, or as far as any milestone in his development. This becomes a matter in which (if we are to consider relevance of measurement), there must be a recognition that tests are an aid to decision; not the instrument of decision and, within the outcome of tests, this has to be put in a perspective that may require a lot of additional non-scorables in the classic sense of scoring -- ideas about what is relevant to an individual in this setting, making this decision.

In other words, our tests should be designed to be commensurate not only to the decision matter, decision information that we need, but it should also be a sample of a kind of behavior that the particular individual or the particular class of individuals is capable of spanning. Until we make the test a reasonable sample of what the individual has learned to do -- that is pertinent to what we want him to be able to do -- we are not performing a rational job of testing. We are just using some convenient set of questions that have been validated against a wisp of population.

PROGRAM CHAIRMAN KELSEY: It is 4:00 o'clock. If there are any of you who wish to stay and continue interaction, please feel free to do so. I think this would be a good note to close on. Thank you very much. Thank you, panelists, for a very stirring discussion.

(Whereupon the session was recessed at 4:00 o'clock p.m.)

- - -

THURSDAY EVENING SESSION

Session One

October 29, 1970

- - -

The Admissions and Admissions Testing Panel Discussion and Workshop was held at 7:30 p.m. on Thursday, October 29, 1970, in the Madison Room of the Hotel Roosevelt. Margaret T. Corey, Director of the Division of Admissions, Testing, and Counseling for Educational Records Bureau, served as hostess. Program Chairman for the evening was Walter W. Birge, Headmaster of The Town School, New York City. The two panelists were The Reverend Canon Harold R. Landon, Headmaster of the Cathedral School, also in New York City, and Paul G. Sanderson, Jr., Assistant Headmaster at Suffield Academy, Suffield, Connecticut. The following is a synthesis of the presentations and discussions that occurred during the evening.

Mr. Birge, speaking first, said the role of the admissions director is to build a school population in accordance with the philosophy of the school. This presents special problems at this time as schools are redefining themselves and their populations. Still, the admissions director must try to balance the student body, Mr. Birge said.

The admissions office can no longer look at test scores and come up with easy answers, he continued, but must also consider many other factors. Tests have fallen into disrepute in some ways, but until other measures which would indicate the chances of a child's success in a particular school are developed, tests will have to be continued.

At the moment, schools seem more enthusiastic about a child who does one thing well, instead of being well rounded in all areas. As the number of applications have decreased in some parts of the country, the profile of the typical youngster which a school is seeking has undergone some changes. In New York City, however, this decline in applications has not yet been noticed, Mr. Birge said. Vast numbers of candidates are still applying for limited numbers of openings.

Emphasis is placed on the role of the admissions office in balancing the socioeconomic mix of a school, Mr. Birge concluded, keeping in mind the availability of scholarship funds as well as other factors. Throughout the entire admissions process it is the role of the admissions director to keep the system humanized and to prevent overemphasis on numbers or mechanical measures of ability.

Mr. Sanderson spoke on the role of the Secondary School Admissions Test in the admissions process.

The Secondary School Admissions Test, the "SSAT," he said, measures both aptitude and achievement for entrance into secondary schools. The most recent development in this field is the availability of a test for entrance into fifth and sixth grade. This was prepared in answer to the criticism that the SSAT was not suitable for such a wide range as it was serving. Although the number of boarding school applicants at fifth and sixth grade is not great, the SSAT Board felt it wise to offer a measure below their former level, according to Mr. Sanderson.

There is awareness that the SSAT has special problems in assessing the academic status of foreign students, he said, as well as students from lower socio-economic groups. It is true that children from upper middle-class families score better, yet some measure of innate ability is needed. The Wechsler tests are considerably not fair, it was agreed.

The scores available on the SSAT are changing, but they are still based on candidates for independent schools rather than those admitted, since the schools do not set up enrolled norms. For this reason, Mr. Sanderson said, percentiles are not necessarily completely valid indicators.

The greatest single indicator of success, the panel members felt, is a combination of admissions testing, previous performance, and the recommendations of prior schools.

Canon Landon spoke last on the subject of the many factors that must be borne in mind during the admissions testing process. He spoke also of the specifics of admissions testing as headmaster of a school with a 25 percent non-white population.

It is important to keep in mind that testing is only one element in the process of evaluation. If testing judges and derogates, it dehumanizes, Canon Landon said; therefore, testing should not be the full factor in choosing children. It is important for professionals to remember, he emphasized, that there is no such thing as a built-in I.Q., that a child's ability is constantly in flux.

The problems of minority admissions are very much in the forefront of our minds, Canon Landon said. It is important to consider testing as only one kind of ability measurement. He reiterated that admissions directors must also consider academic background and performance in prior schools when considering a candidate. Other talents, such as performance in the community, should also be assessed. Above all, a personal interview with each child is especially important, Canon Landon emphasized.

Relating his topic to the experience within his own school, Canon Landon said that it has been his experience that standardized achievement tests and individual test instruments are not completely suitable for minority group children. Children from deprived backgrounds will score lower. "We need to develop new instruments for these children," he said, "and discover other ways to determine their potential." With a determination to meet the needs of these children, as well as the usual independent school candidates, we must see that admissions testing is not a way to keep low the numbers of these children which we serve.

Mrs. Corey gave a summation of the three preceding speakers. In addition, she also spoke in her capacity as director of Educational Records Bureau's Division of Admissions, Testing, and Counseling, regarding the specifics of ERB's Admission Testing Program.

The Bureau's Admissions Testing Program provides a standard testing program rather than tests at a number of schools. It is designed to minimize the pressures on children who apply to more than one school, Mrs. Corey said. In every case a measure of aptitude, as well as achievement in reading and mathematics, is offered. The only exception is at the preschool level where these measures would not have any validity. Mrs. Corey explained that ERB's method of sending a testing specialist right into the nursery school to test young children on their own premises is one of several arrangements designed to minimize pressures on these youngsters. In this way, she said, a child can be tested on any day that his teacher describes him as being most receptive.

Educational Records Bureau, Mrs. Corey continued, has done a statistical analysis of the use of the Wechsler Preschool and Primary Scale of Intelligence. The data discovered from this analysis is available in a paper printed by the Bureau.

Following the presentations by the panel members, there were

questions from the floor and a general sharing of experiences. The audience composition was quite evenly distributed between heads of schools, admissions officers, and representatives of lower and secondary schools.

It was generally agreed that the give-and-take of the question and answer period met well the timely need to capitalize on participants' experiences. The discussion offered solutions to the many problems confronting the admission officer and headmaster in today's schools.

- - -

THURSDAY EVENING SESSION

Session Two

October 29, 1970

- - -

The Thursday evening session of the Educational Records Bureau conference convened in the Terrace Suite of the Hotel Roosevelt, New York, N.Y., October 29, 1970, at 7:30 p.m., with Harry J. Clawar, chairman, presiding.

The purpose of this session is to review some elementary, but often overlooked, test interpretation concepts. We shall deal with three common misuses of the percentile rank, namely, use of the percentile rank to describe group performance on the basis of pupil distribution, use of the percentile rank to compare test performances based on different norm groups, and use of percentile ranks to make judgments concerning growth or change.

1. Use of the percentile rank to describe group performance on the basis of pupil distribution

Error number one can be illustrated by referring to the data in Table 1. Here we find norm tables for independent school pupils and independent school class medians on the vocabulary section of the Cooperative English Test.

Table 1

Percentile Ranks Corresponding to Pupil Converted Scores and Class Medians on the Vocabulary Part of the Cooperative English Test, Form 2A, for Grade 9 Independent School Pupils Tested April, 1965.

%ILE Rank	Class Medians N=142	Pupil Scores N=4039
99	169-172	176-178
98	168	175
97		174
96	167	173
95		
94	166	172
93		
92	165	171
91		
90	164	170
89		
88		
87		169
86	163	
85		
84		168
83		
82		
81	162	
80		167
79		
78		
77		166
76		
75	161	
74		165
73		
72		
71		164
70		
69		
68		163
67	160	
66		
65		162
64		
63		
62		
61		161
60		
59	159	
58		
57		160
56		
55		
54		
53		
52		159
51		

%ILE Rank	Class Medians	Pupil Scores
50	158	
49		
48		158
47		
46		
45		
44		
43		157
42	157	
41		
40		
39		156
38		
37		
36		
35	156	
34		155
33		
32		
31		
30		154
29		
28	155	
27		
26		153
25		
24		
23		
22	154	152
21		
20		
19		
18		151
17	153	
16		
15		150
14		
13		
12	152	149
11		
10		148
9	151	
8		147
7		146
6	150	
5		145
4	149	143-144
3		141-142
2	148	139-140
1	147	123-138

%ILE Rank	Class Medians	Pupil Scores
52	159	
51		
1	147	

Let us take two hypothetical class medians and see what differences in interpretation are produced by looking up percentile ranks in the two norm tables.

The first class earns a median of 164. In the pupil distribution, this performance corresponds to a percentile rank of 71. This performance appears quite good. But, what happens if we look up the percentile rank in the appropriate table, the class-median norm table? A percentile rank of 90 is arrived at. So the actual performance in terms of other classes is much more outstanding than we would have concluded from the use of the pupil norm table.

The second class earns a median of 152. In the pupil distribution, this corresponds to a percentile rank of 22. This performance seems to be quite a bit below average. Again, we look into the class-median norm table for the correct interpretation. A percentile rank of 12 is arrived at. The actual performance in terms of other classes is much worse than we would have thought by using the pupil norm table.

If we were to continue following the above procedure for several more class medians, we would soon arrive at the following conclusion:

> Looking up percentile ranks for class medians in a pupil norm table (instead of a class-average norm table) results in above-average classes appearing poorer than they really are, and below-average classes appearing better than they really are.

A corollary to the above statement is that, generally speaking, the larger the group unit (e.g., entire grade average instead of class average), the more over- and under-interpretation will result.

2. Use of the percentile rank to compare test performances based on different norm groups.

One of the more prevalent errors in test interpretation is to compare, without carefully considering possible norm differences, percentile ranks from aptitude tests with percentile ranks from achievement tests. In order to illustrate the drawbacks of the above procedure, JSAT scores in ERB member schools for pupils taking grade 8 Latin were collected and distributed. Next, a JSAT norm table was built for this group. It was then possible to compare the percentile rank for JSAT scores in the total ERB population with those from the 8th grade Latin group. Six scores were selected for this purpose (see Table 2).

Table 2

Percentile Ranks for Selected JSAT Scores for 'Norm' and End-of-One-Year Latin Group

Grade 8		
	Percentile Rank	
JSAT	Norm	Latin Group
624	90	88
585	79	78
549	66	60
465	38	23
407	19	10
365	10	4

Notice that, in the upper part of the distribution, the assumption of equal percentile ranks is not far from correct. Scores below 500, however, result in quite different percentile ranks. A student at the 38th percentile rank (a score of 465) in the norm group is only at the 23rd percentile rank in the Latin group. It is by no means appropriate, therefore, to assume that a person at the 38th percentile rank on the JSAT should be at the 38th percentile rank in the 8th grade Latin group (i.e., on the Latin Achievement Test).

Several more of these examples would lead to the conclusion:

> To compare two or more test performances, one must be sure not only that those performances are reported in the same unit (e.g., all in percentile ranks), but that the norm groups for all tests are identical.

3. Use of percentile ranks to make judgments concerning growth or change.

A normal curve (see Figure 1) with the baseline divided into 99 equal units (the standile scale) will aid in the illustration of misuse number 3. The area of the curve between the middle of the distribution (median) and a point which exceeds 5 percent more of the area (55th percentile) is shaded. Looking down at the baseline, one can read off the difference between these two points in terms of our equal unit scale. The median corresponds to a standile of 50, and the 55th percentile to a standile of 53. Thus, improving from the 50th percentile to the 55th percentile is a gain of 3 (53-50) equal units.

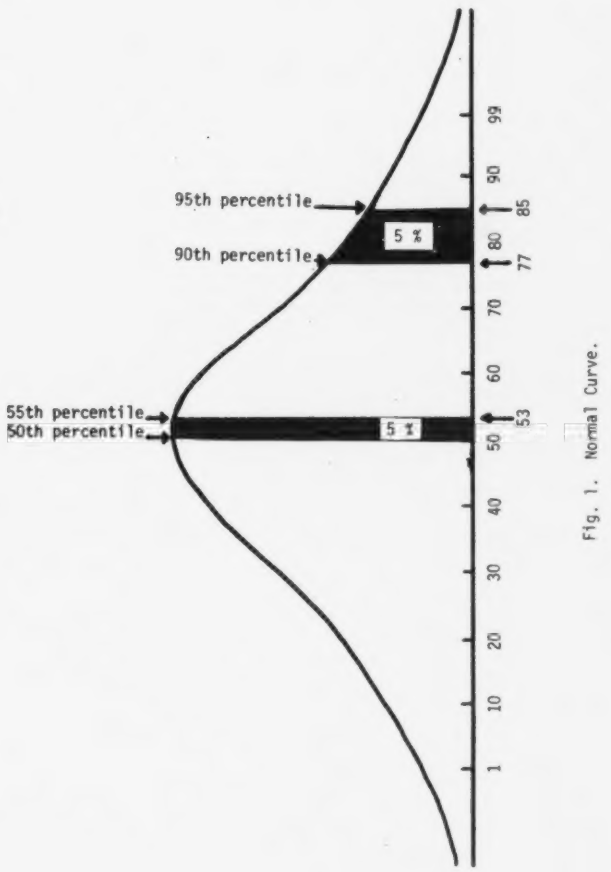


Fig. 1. Normal Curve.

Moving to the right in our curve, we find a second shaded area. This area also represented 5 percent of the cases in distribution. It is, however, the 5 percent between the 90th percentile and the 95th percentile. Look at the baseline and notice the difference on our equal units scale! The gain from the 90th percentile to the 95th percentile is really 8 equal units (85-77). Changing 5 percentile ranks at this point in the distribution results in a gain almost three times as large (8 ÷ 3) as the 5 percentile rank changes from 50-55.

One can see by referring to the normal curve that, as one moves toward the extremes, one finds fewer and fewer people performing at any given point (i.e., the height of the curve becomes lower and lower). It follows that to get a given size percentage slice (i.e., area under the curve) of the distribution, one would have larger and larger distances between the beginning and end of the slice on the baseline of the curve.

It is obviously true, therefore, that a gain of a given number of percentile ranks differs in meaning as a function of where in the distribution that gain takes place. A given size percentile rank gain is much more important at the extremes of the distribution than at its center.

Since a given number of percentile ranks difference has such varying meaning, as a function of the place in the distribution where the difference is observed, it is advisable to compute and compare gains based on an equal unit scare (e.g., standile).

A workshop for ERB members immediately followed the above presentation by Dr. Clawar.

- - -

FRIDAY MORNING SESSION

October 30, 1970

- - -

The Friday morning session of the Thirty-Fifth Annual Educational Conference of the Educational Records Bureau convened in the Grand Ballroom of the Hotel Roosevelt, New York, N.Y., October 30, 1970, and was called to order at 9:05 o'clock a.m. by Mr. Ruchard A. Schlegel, Host Chairman.

HOST CHAIRMAN RICHARD A. SCHLEGEL: Good morning. I am Richard Schlegel, Headmaster of the Detroit Country Day School. It is an honor to greet you and welcome you this morning and briefly introduce the panel.

I have no expertise in this subject, but I have great concern. I think we have the most political topic of the conference -- a controversial topic. I find the patrons of independent schools very concerned about that for which they are spending money and wishing for some type of objective evaluation they can understand, which might go under the heading of "accountability."

I think there is in the public sector a concern for increasing budgets in schools and, along with this, a growing concern for accountability. So, this topic of accountability can suffer from the twin dangers of neglect and abuse.

And we as school administrators and teachers are concerned equally about abuse and neglect. Let me briefly introduce the panelists, beginning with our Chairman, Mr. Robert E. Stake. He is the Associate Director, Center for Instructional Research and Curriculum Evaluation at the University of Illinois. Bob Stake will take over from me as soon as I introduce the rest of our panel.

To my immediate left is George Stern. George is president of the Behavioral Research Laboratories, and was formerly executive vice-president in charge of their financial operations.

To George's left is Don Emery. Don has given us a bit of a scoop. He has just been named Executive Director of the National Reading Center in Washington, D.C. We congratulate you, Don. This is a very important responsibility and we can add this to the rest of the information which I think all of you have on these little sheets.

Finally, at the end of the table is Gary Joselyn. He is School Test Consultant for the Minnesota State Testing Programs, University of Minnesota.

Bob, will you take over from here.

PROGRAM CHAIRMAN ROBERT E. STAKE: I am pleased with the fine turnout this morning for what should be a good hour together. The plan of the day reads something like "us for awhile, then you." We have several ideas we would like to put before you.

Some might see us as a contentious lot. We may stir up a few feelings before we turn the meeting over to you for your reactions and questions. The topic is a broad one, and we are going to narrow it. The topic is "Educational Accountability and the Measurement Task."

I will say a few words about accountability in the large sense, and then we will narrow it to a particular kind of accountability. As a mnemonic device I look to the A,B,C, and D of accountability, something called an Audit, something having to do with Behavior, something having to do with the Curriculum, and something having to do with the Decision processes.

In each of these four areas I see us in the schools as accountable. Accountable to different audiences, to different standard criteria, to our students, to our teachers, to our patrons, to our communities, to our nation, to ourselves. With regard to the Audit, I see us as having practiced reasonable accountability for quite some time: Financial accountability -- legal accountability -- the safety codes -- insurance -- and moral accountability. We are careful whom we hire to be a member of our staff. We watch pretty carefully the moral dimensions of our institutions.

The second area, of Behavior accountability, has to do with student's learning, conduct, attitudes, what sorts of changes are wrought in our children. That is the area we will talk most about this morning.

There is, of course, also accountability as far as Curriculum is concerned. The courses we teach, extra-curricular courses have a certain purpose, a set of purposes, content that needs to undergo continuing review to keep the curriculum in tune. We need to review teaching processes. We need to check continuously to see if we are providing an opportunity for esthetic experiences within the school. This is part of

accountability, of course.

The fourth area, Decision processes, is something to which systems analysts and operations people keep reminding us to pay attention. Not only do we have to make decisions, but we also have to monitor decisions to see if we have internal checks and balances working in the classroom, in the office, and elsewhere. We need to see that our intended values are operationalized. We want to consider the school as a social and political institution -- for it is -- and whether or not it is an effective institution, whether or not it is contributing to the political life of our communities and nation. This we want to check on. All these things have to do with the accountability of our school.

This morning we will consider primarily the problems of measuring achievement under performance contracting. We will consider whether or not there is merit in the many ideas of specifying terminal behaviors, what sort of final performance we want from our students, and the payment of a bonus to various parties who might be responsible for getting that performance.

Before introducing each of the panel members and asking for, perhaps, a ten-minute statement about his viewpoint, his position, what he feels is important, I am going to ask each to identify one principal idea, one thing that they would like us most to keep in mind as we consider this topic, the measurement task in performance contracting.

George, let me start with you. How about one main thought from you first?

MR. GEORGE H. STERN: As a researcher I would suggest that you keep in mind the possibility that you can develop ways of explaining the results of education to the public in a way that they understand, without doing violence to the trust or accuracy or description of the results. Very often results are explained in ways that the average person cannot understand, and I believe that is one of the reasons for the kind of backlash which has led to performance contracting. I would think that a great deal of attention should be paid by the evaluation profession to this particular problem.

PROGRAM CHAIRMAN STAKE: Thanks, George. Don, how about you?

MR. DONALD G. EMERY: For an opening thought, I would like to put a big question mark behind performance contracting by an outside group brought in to see what performance either is achieved or more properly measured. We have a performance contract in the first place in the appointment of the teacher and a commitment to salary. If we need a new kind of performance contract -- that is what we are saying when we go via the Texarkana or other routes -- then I think, rather than embrace that particular concept so wholeheartedly and rapidly, we had better be sure that we cannot get the performance done under the original contract we thought we had.

PROGRAM CHAIRMAN STAKE: Okay. Gary, how about your "thought for the day?"

MR. E. GARY JOSELYN: I think the accurate determination of the level of change in educational achievement is crucial to the basic concept of performance contract, and that is financial reward based on students' performance. It seems to me that most performance contracts, so far, have assumed reliability and precision of educational measurement which may not really exist at this time. If it turns out that we cannot, in fact, measure performance with the accuracy assumed and implied in the usual contract, then the entire model of performance contract, I believe, is called into serious question.

PROGRAM CHAIRMAN STAKE: Very good. With these ideas in mind, let us deal with some specifics. I am going to ask George Stern to speak at some length on basic reasons why many school districts are considering performance contracting now, whether or not this effort to look to outside help is a condition of failure for those of us who have been teaching in and running schools. I am hoping that he will consider what a child is likely to get more of, and what he is likely to get less of, when his school district writes a substantially large performance contract. George is president of Behavioral Research Laboratories. They have at least two major contracts for performance contracting and they have consulted many others. They have conferred with the Federal and state officers with regard to the ways that performance contracting may be carried out. I think we have here a most competent spokesman for the performance contracting movement.

MR. STERN: First of all, I would like to say that I am, by no means, the appointed spokesman for the industry. As a matter of fact, I think that BRL has been the advocate for the industry and probably will continue to be.

The question that I think is in everyone's mind about performance contracting (certainly in mine), is: Is it a gimmick? Is it something that the business community is foisting on the educational field in order to make a few quick bucks? And I have to say that might be the case. I do not know what everybody thinks in the business field. I do know that a number of individuals have formed companies in a big hurry to get in on the act and I also know that individuals who are doing performance contracting sometimes do and sometimes do not know what they are doing in the basic field for which they are contracting. And that means that you have to separate the wheat from the chaff and I think that is mainly your job because as evaluators of education you have to be accountable for whatever your role is in the process.

Now, that being said, I think it is also fair to say there is wheat and chaff in everything, and so there is wheat and chaff in the educational field. Certainly, you know that is true in every area of human endeavor and the fact that there are variations of one sort or another in the performance contracting field does not necessarily mean that performance contracting itself has to go out with the bathwater.

And so I think the questions that have been posed here by Bob are extremely appropriate. Why have we gotten into this thing at all? Why has education suddenly been bombarded with a host of demands (I think that is fair to say) that performance contract be entered into so that certain things will be more finite?

Well, I think one of the reasons is implied in the question. Things are, in a way, not finite in education. It is well and good to say that education is partially immeasurable because of social and political objectives. Nevertheless, there comes a point where one has to be able to say, with some degree of precision, what is going on at that point. Unfortunately, this comes in the midst of the political upheaval that so much of the country is going through now, where a number of individuals are not being served by education. As you well know, a number are. Individuals who are not being served by education are the poor, the black, the brown, the Indians, and possibly other groups that for some reason or another have not been able to get in on the act.

Unfortunately, education seems to get blamed for everything because everyone recognizes how important it is. This is unfair, but still education has the limelight, partially because of the amount of funding devoted to it. As a result, education has become, in a way, the whipping boy for the failures of society.

When a taxpayer is required to, let us say, divest himself of what he considers a large portion of his fairly gotten gain, he wants to know why. On the one hand, maybe it is a little easier when the money is going to Vietnam than when it is going to his public schools. That is a political problem, but there is no question that over the last three or four years public support for public education has become at least precarious, and I think it is also fair to say that one of the causes has not been the failure of education -- although I think things could be said about that -- but it has been because of the ever-increasing costs that seem to have no end. Now we are paying teachers fairly and now we are paying other individuals associated with the school systems fairly, or almost fairly.

Now that school systems are in the public press all the time for reasons that frequently do not have anything to do with education, individuals who are paying more and more for that public education are wondering why they seem to be getting less and less. Whether they are or not, I think, is a very complex question, but they think they are. And they feel that if they are to pay more and more they want some tangible evidence of what they are paying for. They want to reap some results.

Now, as well as that, they read that there is an increasing crime rate, an increasing rate of drug abuse. There is racial turmoil which is manifested in many ways. There is a great deal of dissatisfaction within the profession, which manifests itself in various ways from teacher strikes to dissatisfaction with the particular goal of education in any given school system. The original individual who, after all, is the person the educational system serves, is becoming a little on edge and in the course of his dissatisfaction he is grasping for something specific.

Now, to return to what I said in the first place: it is very difficult to get specific, but in a way we all have to steel ourselves to get specific. And it seems to me that one of the things the evaluation profession had better do is to think of ways to explain to the public -- the public that is responsible for demands on educational systems that have led to performance contracting -- what it is they are getting from the educational system, not just in vague terms, but in terms of specific, basic skills that apply to the population with whom the public is concerned. Of course, the populations that seem to be the most vocal are the populations that do not feel they are being served by the educational systems, and those are the populations to which you should address yourself in your reports. I can not tell you how, but I suggest that it is your responsibility to find ways to do this.

Is it an admission of failure on the part of the educational

system that it has to turn to such things as performance contracting? In my opinion, the answer to that is no. I think that the reason it has seemed to be an admission of failure is because an outside agency has begun to enter the field and outside resources are being applied to the task that was formerly considered strictly the role of educators.

Educators, I think, like all professionals really do not like to be intruded upon in any way. It would seem to me that educators should be willing to turn to whatever technology or ideas or pressures are available as agents for change in the entire country or in the world.

And the fact that those particular resources happen to be in the areas of business that are recognized as "dirty words" in many circles, is something that I think educators should put in suspension until they can evaluate the contribution they are getting.

It turns out that some of those educators who have turned to performance contracting have turned to it not as an admission of failure but as an expression of their determination to change the particular patterns that have existed for certain children in their school systems.

If it turned out to be a failure or an admission of failure every time we did something new, then you can be darned sure we would never do anything new. I would suggest to you it is not an admission of failure on the part of any individual system to attempt performance contracting, though I think we could quibble with particular performance contracts, but is instead an expression of determination to change the particular problems or to solve particular problems in the middle of which the educational system has found itself. And I don't think that is an admission of failure at all.

Bob has asked what the child will get more of and what he will get less of. Hopefully, in the better performance contracts the child will get more of the basic skills. Of course, I mean in a performance contract that works, because the basic skills are more measurable. Also, they happen to be, I think, more important, at least as a starter. But the reason I think the child will get more of the basic skills is that it is much easier to tell whether the child has advanced or not in the basic skills. Again, I am afraid that is your problem.

The thing the child may get less of is the sort of arrangement in a school which puts him in a holding pattern simply because he is either obstreperous or irregular or undisciplined or strange or unconventional or something else which, for some reason, has him tagged as a problem.

In the course of a performance contract that works, a child who is a problem must be involved (and in my opinion this is where educational systems should work). In general, there is no way that a performance contractor or any person responsible for education can cop out of dealing with a particular child that he has in front of him and, therefore, I think it very likely that the performance contractor whose rewards are going to be determined according to the results achieved by this particular child will be forced to change the ways of dealing with that child so that they will indeed cause him to achieve results.

I think this has more to do with determination than with technology. And this, I think, is a major point that must be emphasized in performance contracting. In my opinion, the primary failure of the educational system, if there has been one, has been to grit its teeth and to change its ways so as to deal with what it considers problems of the individual child, and the many individual children, particularly from the disadvantaged areas, whom it is required to teach.

There obviously have been methods and ways of dealing with children, but it seems to me that those methods were ways of avoiding the real problem or the real task of bringing about achievement and, instead, have been reasons for excusing the fact when achievement does not take place.

This brings us right back to the beginning of why performance contracting has become such a national fad. And I think it poses the most important problem that educators have before them today -- that is whether they, themselves, can change their approach to children so that they feel obligated to produce results, even with children who have been called unteachable. The premise of performance contracting is that there are no unteachable children, and it seems to me that what the public is saying when it demands performance contracting is that there are no unteachable children, particularly our children, particularly my kid.

My kid is not unteachable. You are the educators; so you think of a way to teach him. I realize that is not easy, but I suggest to you that educators have not really applied themselves to the problem as thoroughly as they could, and that is one of the major reasons why performance contracting is taking place.

It is easy enough, I think, to be deluded by financial considerations. As a matter of fact, financial considerations are, in my opinion, the reason performance contracting will be gone in two to three years. But the major thrust of performance contracting ought to be as an agent of change for the present educational systems so that there are ways of producing results within the present educational systems, and with the present teaching staff, by means of advanced techniques. By that I do not necessarily mean scientific techniques, so that the public educational system, as a public educational system, can do its own job to satisfy the consumers who, after all, are the parents and the children.

PROGRAM CHAIRMAN STAKE: Thank you very much, George. That was a farsighted and well-reasoned statement. That was not very contentious so we may not have as much argument as I thought.

Don Emery, is there any problem of agreement between the local faculty and the performance contractor as to what criterion tests should be? What are the roles for the teacher and the administrator in the performance contract?

MR. EMERY: I think we are likely to echo certain things back and forth among us. Yes, I think there is a lot of ground for disagreement and some for agreement. I would like to speak a little generally to the proposition first.

Let me say that I think the concept of accountability and the practice of performance contracting comprise probably one of the best things that has happened to us. Even though I have a number of questions about this, I believe by and large (regardless of how this may appear five years from now), the fact that we are now caught up in these dual propositions will be a healthy thing for education.

George commented about some of the circumstances that have triggered the popular notion about accountability and performance contracting as one way to resolve the dilemma. Certainly, we have to be concerned with increasing costs, which are attributable to the fact of inflation.

When the taxpayer looks at a series of tax bills or tuition bills over five or six years, he says, "My God, what is happening to this thing?" And he forgets what is happening to his salary. He looks at the cost of education as an isolated item in comparing a series of tuition and tax bills and comes up with a slightly disordered impression beyond what the fact is.

The increasing level of frustration with the results of education in all of society, the best of the independent schools, the best of our affluent suburban schools as well as the worst of Appalachian or urban or ghetto schools. Nobody is doing nearly as good a job as is needed in this nation or could be done and the taxpayer knows it. Parents know it. You know it. And I know it. And we say, "Let us go on with it, then."

Why all these frustrations? Kids are compounding the problem with dress and protest and drugs so that the whole climate is very difficult. The frustrations are present at almost any dimension you want to select as contributing to a growing demand that we be more accountable in what we are supposed to be doing, which is not an unreasonable expectation.

The public and the parents thought that was what we were there for. Society created us for that reason. We are the performance contractors for the society at large, created by legislators to do this job. Now we, in turn, are raising questions -- maybe somebody else has to do the job or part of it -- and we should subcontract because we are the prime contractor of society through the state legislatures.

In addition to costs and frustrations, there is an overt and definite demand by the teachers as a professional group that they have a much larger piece of the action in terms of decision-making and control, versus the Board of Education or Board of Control or Board of Trustees -- the administration, whatever it may be.

I believe when you take those three matters together -- greater costs, greater frustrations, and debate over who is in charge here -- you are bound to get this kind of question of what is going on here and how well is it being done in the midst of all this business, and I think a very natural suggestion would be: "Let us hire somebody else to do part of this."

Or, an easy solution in the mind of the Board of Control might be: "Maybe we can get somebody else to do this if you boys can't do it. It isn't so far out after all, because we do hire architects, lawyers, and all kinds of special people to come and do temporary jobs for us, but we have not hired anybody to come and do the basic job for which the school exists."

This, in effect, is what performance contracting is about,

to some degree. Now, we haven't been upset about bringing aides into school to do some of the so-called chore work and set teachers free to do more of what teachers should do. That concept has not upset us. We have gladly accepted the various forms in which media can be used. If teachers use films as part of the teaching-learning process -- that has not upset us. But now, when it is proposed that somebody else literally do the job we felt teachers were contracted to do in the first place, that should be an upsetting proposition, most of all to the teachers.

If we say that somebody else can come in and teach reading better or teach mathematical skills or understanding better than the teacher we hired, then something certainly is wrong if it can be done that way on any grand scale. For years and years and years parents and taxpayers smiled at teachers and were nice to teachers and patted them on the head because they had a good thing going.

Teachers were underpaid for a long, long time. We are much closer to a pariety relation in the economy now, and when the tax bill is at the level established as pariety professional salary, then the person who provides those funds is entirely correct in wanting to be surer now if you really have to pay, and he has to pay now, and will have to continue to pay, he is entirely correct in asking, "What am I getting for it? Tell me why you say that, or get a piece of paper and show me. I am tired of having people say nice things and go away."

The real question then, I think, is: Are we willing to redesign our expectations in the process of getting the results we want? Accountability is not a concept foreign to educators or to the profession. It is exactly appropriate to our profession. It is a preciseness, a completeness and a sureness of accountability that is being sought.

To the best of my knowledge performance contracting is not reaching into fields of understanding and behavior and attitude. They would like to perform in the nice, clean, neat area of skill demonstrations. You either can read or or you cannot at a certain level. You either can compute or you cannot at a certain level. You either can type or you cannot at a certain rate. In some ways the easiest part of the job might be contracted out, leaving the profession still with the hardest part of the job and maybe what they should have as their unique area of responsibility, are the behaviors, attitudes, values and understandings.

I pulled out a few of our elementary teachers' schedule cards a couple of days ago to try to generate a little more feeling for what remarks I might make here. And I was struck by what we are expecting a teacher to do in a primary grade, and what advantage I think a performance contractor has in the situation.

What the record reflected was that, historically, we have been willing to take on every good goal, object and idea from pressure groups of all kinds and we are trying to do far, far too much. We have not stretched the school day, but, oh, the things we are trying to do in that school day. If I were a performance contractor, I would say, "Okay, that is exactly what I am going to do. You are not going to expect anything else. Right?"

"Right."

"You are not going to interrupt me. You are going to give me the teaching machines and technology that are part of a contract which states that I am to have the tools I really need to teach?"

"Sure."

"And you are quite satisfied that I can employ a series of tangible rewards immediately?"

"Yes."

"You are not going to give me too many kids?"

"No."

Okay, let us do that <u>inside the school system</u>. Let us quit asking teachers to do everything under the sun -- teaching and other things, too. Let us quit interrupting them. Let us give them more tools. Let us help them in the measuring, and maybe the rewards will be more immediate for both the learner and the teacher.

I think very, very important principles are involved in executing performance contracts that we are violating in our own schools. These violations help to provide the opportunity for performance contracting to come into the picture. I would raise a very serious question as to whether a substitute procedure is really needed.

On the other hand, if we could find cheaper ways to sub-contract part of our job and to free the teacher for more difficult andmore skillful contributions to the goals of achievement of the school -- that would be highly desirable, too.

I think faculty should be alarmed and concerned about performance contracts. They should cause us all to rethink our purposes in education. Thank you.

PROGRAM CHAIRMAN STAKE: Thank you, Don.

We are beset with problems regarding the ability to discern the effects of our teaching programs. Gary Joselyn raised that issue already in his opening comment. I hope he is going to deal particularly with the question of how about this "teaching and test" business?

MR. JOSELYN: Let me start by relating an actual personal experience which, I believe, illustrates many of the philosophical and technical problems that confront us with performance contracting. The very next morning after Bob called me about this panel I received a call from a school psychologist in a medium-sized town in Western Minnesota. His question was: "Are grade-equivalent scores available for the high school level Standard Achievement Test?" After saying they were not, I asked him why he wanted them, and his explanation went something like this: several of the school's teachers had gone to Texarkana to see the performance contract in operation there.

Upon returning, the math teachers went to the administration and school board and said that the idea of pay being based in part on students' performance seemed like a good one to them. But rather than hiring outsiders to come in and raise the achievement level in the high school, why not instead give them, the math teachers, a bonus for causing their pupils to achieve "above average."

The proposition they presented, and which the board apparently accepted, was that the math teachers would receive a bonus for every pupil who "grew" in mathematics at least one and a half grade levels during the year.

The bonus would increase for each additional half-grade level of growth. The school psychologist had been asked by the superintendent to find an achievement test to measure the students' growth and to award the bonuses. So, one brand of performance contracting has even reached rural Minnesota and if we could not just get Harcourt-Brace to come up with some grade-equivalent scores for the high school SAT everything would be rosy.

In civilian life I am a member of the school board of a large suburban Minneapolis school district, and my first "gut" reaction was, "Damn it, if you guys are able to do this, why aren't you doing it already?"

I digressed from our task of considering the measurement implications of performance contracting, but I hope this incident helps to illustrate some of these problems. The previous speakers have dealt mostly with what I would call philosophical aspects of this issue and I would like to talk about those, too, but Bob has asked me questions relating to measurement.

One of the questions Bob asked in a letter he sent to us was, "I would like to know what differences we may expect from test results if the tests are designed to measure exactly what the curriculum teaches or only what it teaches in general? Perhaps we can envision these two cases as the opposite ends of a continuum. Developing tests to measure exactly what the curriculum teaches, carried to the extreme, would probably have "exactly what the curriculum teaches," defined simply in terms of a pool of test items -- it would be difficult to be more exact than that. It would then be the task of the contractor to teach the answers to the items and to measure how well he had done. We would simply administer the items to the students at the conclusion of the contract.

Taking the other end of the continuum to the extreme, testing generally what the curriculum teaches, we would hire the contractor to teach reading or mathematics without purpose specifications of the curriculum content. The test could then be any reading or math test the evaluators might choose to administer, or the school could simply make a subjective judgment as to whether or not to pay the contractor.

Now, these two extreme situations are absurd, but they do, I believe, represent the logical extremes of the two situations. In every contract the contractor and the school will eventually have to settle upon a position somewhere between the two extremes. And there will be both positive and negative pay-offs as we move off from one extreme to the other. The first situation, where the curriculum is defined by specific test items, is the easiest and "fairest" to the contractor. His task is clearly defined, the goals are obvious, and at the end it will be quite clear to everyone whether or not the contract has been fulfilled.

However, schools are not likely to settle for this approach. They will argue that "there is more to reading than can be

defined by a pool of test items." While "fair" to the contractor, this condition may not be "fair" to schools seeking a good general education for their students. They may also feel that this model contains too much temptation for the contractor to concentrate on tests and take shortcuts.

At the other end, where the curriculum content has not been specified at all, and the school could use any measure it chooses to determine outcomes, the conditions are not very fair to the contractor. How can he proceed with the teaching task if he does not know in advance what the school will judge to be important and what kind of a determination it will make as to whether or not he has fulfilled his part of the bargain?

The contractor must know, in advance, what the ground rules are and will want them spelled out in as much detail as possible. My point is that the testing aspects of every performance contract will be a compromise, a trade-off between very specific, narrowly defined and easily measured outcomes and general, broad outcomes which seem to come closer to capturing the true essence of the subject area, but which are much more difficult to measure.

A related difficulty, as I see it, is the question of what the measurement payoff is to be based upon. Some contracts are based upon a guarantee to bring student achievement scores up to a specified level, like, for example, the EDL contract with San Diego which, according to the June 1970 "Phi Delta Kappan" guarantees, "that students will achieve 25 percent closer to the city norm during the first year, 50 percent closer during the second, and at the same level during the third."

I could not help but wonder what was happening to the city norm during these three years. Other contracts, like that of my Minnesota math teachers, are based upon gains or change scores. If we aspire to bring all pupils up to a prespecified level, what about individual differences?

Certainly, some pupils will already be above the level before the instruction starts, while others will be so far below that it will be impossible to raise them to the payoff level whatever the treatment. If, on the other hand, we base our payoff on gain scores, we run into the tremendous technical problem associated with the measurement of gain. I am not an expert in these statistical considerations but, the way I read the literature, the general agreement by those who are experts seems to be that gain scores are generally useless, if not impossible to get.

Breiter, in the Harris book on the problems in measuring change, states that it is only with regard to problems in measuring change that he has ever heard colleagues admit to having abandoned major research objectives solely because the statistical problems seemed insurmountable.

More recently, Cronbach and Furby concluded that we should not attempt to measure change in most instances. Another concern of mine is the timing of the measurement. For many programs we are not interested so much in what happens immediately at completion as we are in its longer-term effects.

If our goal is "an improved attitude toward learning," for example, our end of the program measures cannot really tell us how that which was learned is retained, applied, or transferred to new, future life situations. This concern was expressed for a less exotic outcome by Elam in the June "Phi Delta Kappan." He was unhappy that the US Office of Education had deleted a clause from the original agreement between the school and Dorsett in Texarkana which would have provided for retesting six months after treatment to determine whether retention rates were equal to those of the average student within the system.

He said, "Thus temporary achievement spurts so familiar to educational researchers -- usually due to all those factors we lump together as the Hawthorne effect -- may fade away without anybody noticing." The question has been asked, "Do we need to worry about 'testing the test'?"

Recently published information about the Texarkana project shows we certainly do. It seems pretty clear that students were taught at least some of the specific items that were included in the final test.

The amount of true contamination is not clear. Estimates run from 6.5 percent (Dorsett) to 100 percent (according to the independent evaluators). By teaching the test I here refer to teaching students the answers to specific items which later appear in the final test. Teaching the content areas which the test measures, of course, may not be bad.

In fact, if we have agreed upon certain desired outcomes for the program and have designed a test to measure those outcomes, we want the instruction to teach for the test. But teaching students the answers to items which sample a particular outcome hardly tells us whether or not the student knows anything about the desired outcome at all.

As long as the entire payoff is based upon the scores on a test, contractors will certainly teach for the test. We are telling them to do just that when we tell them that their reward, if any, will be based upon that test score. Another question Bob asked in the letter was, "What about using standardized tests as criterion measures?"

This is, in my opinion, a terribly inappropriate use for such instruments and a use for which they certainly are not intended. Our present standardized tests are designed to maximize the discrimination between individuals. The difference between students in their knowledge of arithmetic, for example, may be very small in relation to the knowledge gained by all students during the course of instruction.

Yet, standardized tests attempt to magnify those small differences. Items which most everyone misses or gets right are eliminated in the test development process since they show only how the students are similar, not different.

The best items are considered to be those with a 50 percent difficulty level. Thus, the very items which have the most potential for measuring change or status are the ones which are eliminated. The grade-equivalent score is used as the measure of gain in many contracts. That is, the contractor contracts to raise students' scores by one or more grade levels. What could be more beautiful? Here we have a group of students who are tested to be three grade levels behind, and among them a contractor who will guarantee to raise their achievement one or two grade levels or else you don't pay him.

Free enterprise has finally arrived in education. We have free enterprise, but do we have education? Let us look a little closer at what our enterpreneur has contracted to do. On the Arithmetic Computation Test of the SAT, Advanced Battery (Junior high school level), a student answering six of 44 items correctly gets a grade equivalent score of 4.2; if he gives nine answers correctly -- that is three more -- his grade equivalent score is 5.1, just short of a year's "growth."

On the arithmetic problems test of ITBS for grade seven, which is a 32-item test, a student grows approximately one year in grade equivalency for every two and a half additional items answered correctly. This is from a grade level of four where what you would get with five items correct would be a chance score right up through a grade equivalent of 8.0 and he gets that with 15 items right.

So, if the payoff were based on a guarantee of raising the student's grade level one year and this particular standardized test were used as the criterion, the contractor would get paid if, for example, he could raise the student's raw score from seven to nine on a 32-item test. And let me remind you that this item pool will always be known to the contractor.

I don't think standardized tests will work as criteria for performance contracting. They may appear to work for some administrators and school boards and I expect the contractors will like them. But we simply must pay more attention to the measurement aspects of performance contract than has been the case up to this time.

PROGRAM CHAIRMAN STAKE: Let us see if we have a reaction first from any of the panel members to any of the statements. Then, we will take some questions or comments from the floor.

HOST CHAIRMAN SCHLEGEL: I would like to add one thought from the point of view of the independent school, since the panel primarily has been addressing its thoughts to the larger number of our students in public schools, and quite rightly.

We are not free of this concern nor this responsibility in the independent school. I think most of us recognize this. If we do not, I think we are making a grave mistake. The tuition that our parents pay is a legal contract with the school and if any of us in the independent schools think it is not legally binding and cannot be tested, then we are ignorant of the facts.

This, in a sense, is a performance contract, and parents are beginning increasingly to push to have this performance measured. Now, one other brief thought: this was an article published recently in the New York Times, and similar articles have been published regarding the number of students that now are admitted to college from the so-called prep schools as though the criterion of evaluation parents are using is getting into college, whereas the fact is that getting into college no longer provides sufficient criteria for accountability to an independent school. More is wanted, for the high tuition parents are paying. Our costs per child, to our surprise, calculated on the basis of figures received from the state, run, relative to the teacher-student ratio, about the same as public schools.

Now, the criteria for admissions to colleges is simply not sufficient for our parents to pay that kind of money. So, we, too, share in these problems in a very real way.

PROGRAM CHAIRMAN STAKE: Don, or George, how about you?

MR. EMERY: I would like to say that I think in the last set of remarks a very important matter was touched on and that is the necessity for the parent and board to know what is really being done under the performance contract, so usefully illustrated in a couple of items on the Standardized Achievement test. That is not at all the way you measure things when you achieve enough level of progress.

You need to know what kind of measuring is really appropriate under a performance contract before you try to execute one.

MR. STERN: I would agree with that. I would really rather hear from the audience.

PROGRAM CHAIRMAN STAKE: I will take this lady down here.

MEMBER: I just want to inquire whether any measures of ability of students to begin with is involved in this kind of contract. It is one thing to achieve a year or two years' progress with an extremely capable group in comparison with a group that comes less well endowed. How do you allow for that difference.

MR. STERN: Given problems in measurement devices have been well highlighted here this morning. Usually a performance contract does have a preliminary measure of what the children's achievement level is at the point they entered what you refer to as the treatment phase. At least, any performance contract I have seen has made an effort to identify that. Certainly it should -- particularly if there is any sort of interpretation of results being given to a degree of progress. Obviously, you would have to know the starting point in order to tell the differences.

MEMBER: You have to know the ability, not just the achievement level, and they are two different things.

MR. STERN: I think one of the most important things about performance contracting is that it tends to erase the notion of ability levels. I think ability levels have tended to become something that has overshadowed the need of the public schools to deal with all children.

Ability levels, in the first place, are subject to the same levels as achievement levels, and as a result I believe that we have tagged some children -- who simply do not come to school with exactly the same experience level as other children -- as low ability students.

PROGRAM CHAIRMAN STAKE: We have a question in the middle here. Yes, sir.

MEMBER: I want to extend a proposition which would say, suppose for instance the third grade vocabulary was to be taught by a performance contractor and that this third grade vocabulary would be defined as a set of some thousand words, what is to prevent the independent mediator between the contractor and the school district or school from selecting towards the tail end of the term, or the year, some sample from that thousand-word vocabulary bank -- say 50 items -- which will include some adverbs, some nouns, or whatever, and that performance be tied to the performance on that 50-item test.

The ground rules for the way in which performance would be tied to outcomes -- whether a child would get one right, two right, or 50 right -- could be arranged in advance by the contractor in the school district and mediated again by an independent contractor, but the criterion for payment would be independent and would be selected independently of both contractor and school district. And that is a proposition.

MR. JOSELYN: I think, for the particular proposition raised, it would be a very valuable model and would work.

MEMBER: Could it be extended?

MR. JOSELYN: That is where you get in trouble. It seems to me, if the task you assign to the contractor is to teach vocabulary, period. In other words, can we upgrade the students' ability to pick out the correct definitions on multiple choice tests? If so, our domain of achievement is pretty well defined.

But what if you are going to hire him to teach reading or arithmetic? Then, where are you going to get items? In order to define them, it seems to me you need thousands and thousands of items. In your other proposition you said "define for the great vocabularies."

Everybody can get a handle on that. If you go beyond those kinds of very, very specific things, both the contractor and the school will be in a lot of trouble trying to decide how are we going to say whether or not he did what we wanted him to do.

PROGRAM CHAIRMAN STAKE: Yes, sir, way back there.

MEMBER: I wonder if we can't define that very difficult thing just mentioned, what we tell ourselves at the end of the year we have taught our children. How do we know they have made any progress? We must have defined it all these years we have been doing it to one another. And if it is definable, if this is achievement, it must be achievement in some specific manner.

I don't see why it suddenly becomes difficult to say to the contractor or to the teacher in my district, in my school, "Gee whiz, your kids aren't doing as well or don't arrive at this point. You teach them specific kinds of tests, you really want to know specific kinds of things." I don't see why you don't want to do it. I am afraid that, in refusing to separate skills from attitudes, we keep on telling ourselves that we are doing something else in teaching besides teaching what can be learned.

PROGRAM CHAIRMAN STAKE: Do we have a reaction from the panel?

MR. EMERY: I have a feeling we haven't been nearly as discrete as some of us think we are. As to what can be done this year and whether it did in fact get done, this is a function of inadequate planning, staff, administration. This, I think, suggests a very important area as to why we are not more specific about this whole business. I think we could measurably improve.

PROGRAM CHAIRMAN STAKE: May I also react to that question? I think there is a tremendous difference between actuarial measurement and the clinical intuitive measurement of which professional people are capable. We have not been trained to specify the operationalisms of reading ability, the capacity for doing certain things.

We have been trained, I believe, to recognize good and bad reading behavior in a classroom setting. Just because we can recognize things does not mean we can specify them or put them into some sort of a contract form.

I feel confident that we could write contracts whereby reading specialists would spend an hour with each student at the end of the contract period and have a very clear idea of his reading level, not expressed in grade equivalency, not expressed in test scores, but in terms of how good a reader he is in the many dimensions of reading. In trying to put that in contract form, trying to tell somebody else about it, we lack competency.

MR. STERN: Excuse me, are you suggesting we get rid of standardized tests?

PROGRAM CHAIRMAN STAKE: No, I am not suggesting we get rid of standardized tests; I am saying there is a skill in measuring things that is being ignored, for some very practical reasons perhaps. Like Gary, I feel that the use of standardized tests as the sole determinant of reading skill is questionable.

HOST CHAIRMAN SCHLEGEL: We are running late. I would like to add one comment. I think we haven't touched upon the area of community which we once shared in public education. And much of the terminology, even in the skill and non-skill areas, would lack understanding and communication because of the lack of community.

Thank you for your attention.

(Whereupon a short recess was taken from 10:20 until 10:30 a.m.)

- - -

FRIDAY MORNING SESSION

Session Two

October 30, 1970

- - -

The "Individualized Instruction: The Measurement Dilemma" session of the Thirty-Fifth Annual Educational Conference of the Educational Records Bureau convened in the Grand Ballroom of the Hotel Roosevelt, New York, N.Y., Friday morning, October 30, 1970, and was called to order at 10:25 a.m., by Chairman Harry K. Herrick.

CHAIRMAN HARRY K. HERRICK: I am happy to welcome you to this session "Individualized Instruction: The Measurement Dilemma." Dr. Uvaldo Palomares, our speaker, is co-director of the Human Development Training Institute and president of

the Institute for Personal Effectiveness in Children in San Diego, California. Dr. Palomares is on leave from the Department of Counseling and Guidance, San Diego State College. He did his undergraduate work at Chapman College and San Diego State College in California. He completed his Master's and Doctor's program at the University of Southern California in Los Angeles, earning his degree in Educational Psychology and Elementary Administration. He has also done advanced work in research statistics and educational administration and supervision.

Dr. Palomares was born and reared in a Spanish-speaking environment. As a child he traveled a great deal through Arizona and California while his family was engaged in picking crops. His first interests lie in the areas relating to the Mexican-American sociological group and in educational psychology, education of the culturally disadvantaged, compensatory education of the migrant, clinical psychology and early childhood guidance. His background has made him an authority in this area. Among the many positions he holds are: Consultant to the US Commission on Civil Rights, member of the National Advisory Committee on Educational Laboratories and, as of last year, special consultant to the Secretary of Health, Education and Welfare in Washington, D.C.

It is my pleasure to introduce Dr. Palomares.

DR. UVALDO H. PALOMARES: I am not going to give a long pretalk before I start answering questions. I am going to make a few statements about where my thinking is, and an assessment, on a realistic "nitty-gritty" basis, on an "out there, right there, about-to-be-there-tomorrow" type world, and then I am going to start answering your questions.

Okay? It is necessary to hear these statements because it will put you in tune with where I am and then maybe we can go somewhere together. We have all, as administrators, had a fear when we stop to think about it. The fear is something like this: That, sooner or later, if we didn't get hold of ourselves somebody was going to get hold of us, and it is happening; that, if we didn't start getting better at our jobs, somebody was going to make us get better. We wanted to play professional but we allowed ourselves to be put in positions where we couldn't be professional and now we are paying the price for it.

I am involved right now in a movement to create educators -- and particularly people interested in the whole business of measurement -- a movement for whipping us into shape. Let me tell you what I mean by that. How can I put it into succinct words?

Until now, government intervention through the Civil Rights Commission and other, similar branches has been based on the whole business of segregation. Black kids and white kids have not been grouped together because when they are grouped together they learn more than when they are separate.

Until now, the issue has been busing and so on, and so on, and so on. These are not necessarily educational issues. However, there have been other groups than blacks who are very concerned and one has been the Mexican-American. That is what I am, Mexican-American. And we are calling ourselves "Chicanos" now, so we are using our own terminology. I don't want necessarily to get into that unless that is the area you want to explore. The point is, we felt that bad things were being done to us and we wanted to do something about it. Educators and theoreticians, and so forth, and so forth, pretty well knew that we were getting a bad deal through education, but nobody knew how to articulate it or put it together.

In the past two years very dramatic things have been happening that mean a lot for you and will mean a lot more for you in a couple of months. What happened was the way I thought about it and the way other people began thinking about it. We began to say that, if a school district includes a high percentage of any unidentifiable subgroup having specific educational needs, and if that district does not have a specialized educational focus and emphasis to deal with its particular needs, then the school is robbing those children of equal educational opportunities, and there is nothing as guilty of destroying a person as a school that is segregated.

A specific example is a Mexican-American subgroup. Schools are not giving the children equal educational opportunities because they tested them all with the same test, taught them with the same books, treated them all the same. Some of these children could speak very little English, some none at all, and some a mixture of both. Not only did they have linguistic handicaps, but they had a different culture. Yet the same instrumentation was used and the same basis of instruction. And we call this equal educational opportunity.

I and many people like me disagree with this. We say that, so long as a district is not actively moving toward providing educational programs to meet the needs of these children, that district is not offering equal educational opportunity and is, therefore, liable for malpractice to those children.

Now, we don't mind the districts being ignorant, but we do mind their not moving in the right direction. Therefore, because we know the history of keeping kids separate, we can take a district to court for keeping them separate and thereby robbing them of equal educational opportunities. We are robbing them of equal educational opportunities by the type of situation that exists when we have subgroups that differ significantly from the majority population.

Until now, all we had were words, or ideas in the minds of a few people like me who may be considered by their colleagues to be rather excited. About four years ago I did a study in Imperial County in which I took a series of individualized tests like the Stanford, the Binet, and so on, and administered them to prove that tests did not work well. The interpretations of instructions differed radically and, therefore, led to misplacement and so on. I had the results published hoping that this would miraculously cause a change in attitude. Well, nothing happened, and I found out that a guy in Texas had done an identical study two years before I was born. That is a heck of a thing to find out after working on it for a year.

I am saying all of those things to lead up to this: as of three months ago the Civil Rights Commission sent out a mandate that districts having no individualized programs, or programs suitable for subgroups, could be held liable and taken to court because they were not offering equal educational opportunities to children.

Two months from today the Commission will come up with the first two districts who have Puerto Rican children, black children, Mexican-American children, French-English speaking children -- pockets of significantly different kids -- and if those districts do not have a specific program to meet their particular needs or a plan to move in that direction, you people -- you, we, educators -- can be taken to court and sued for not offering equal educational opportunities.

Okay. Now, I will just summarize: I began by saying that I felt very bad because we all had a fear when we thought about these things and the fear was that if we didn't clean house it would be cleaned for us. I must say that I am critically involved with the Federal government in setting up the machinery to help us clean house.

The key target will be the misplacement of children in mentally retarded classes, and then it will spread from there. As a professional, I am shocked at myself for having done that. I have seen these youngsters suffer and become burdens on society. I will stop here and see if you have any questions you want to ask.

MEMBER: My question must, I think, be introduced by a couple of comments. If we were to agree with you that if you have a significant subgroup or certain value or culture that needs protection, let us accept that and say, yes, if you intend somehow to protect those values and those cultures that have a significant impact on society, this is desirable. I think, however, there is a contrasting factor here that seems to have been omitted completely -- and that is the concept of assimilation in a very positive form and one can ask the question, "at which point is assimilation desirable?" and, since we are dealing here essentially with subjective philosophies and values, I think the question should be raised.

If that question is raised, then I think we have a scientific problem as to how significant a group becomes and under what conditions. In other words, would, say, two or three individuals out of ten thus be significant? I think the answer to that would probably be no.

Would 30 percent be significant? A quick answer would be yes. However, it seems to me the real challenge is at which point and who decides, if not society, and how you constitute it as to what number is significant in order to justify legal action to protect the values of that particular culture in that society?

DR. PALOMARES: You know, I may try to repeat your question to see if I heard you correctly. I have been forming my answer so rapidly that I never paid attention.

I heard you ask me about my view of assimilation, tending to be the way things happen in this country, and kids having to fit in, and succeed, and survive within this culture. How does this stand up against the statements I made prior to this? When does assimilation start taking place? When does a child stop living in his own culture? That is one thing I heard you say. The other one is: when do we start considering a group significantly large enough so that we begin introducing special programs to meet its needs at tremendous expense in districts where such changes would diminish educational systems for other kids? When does that happen? You said something else, but I forgot it. What was it?

MEMBER: I think you did extremely well in getting the sense of my question. The significance is correctly stated. The assimilation is a little bit misstated. The question I really raised is: what is your notion as to what conditions might render assimilation a desirable aspect of society? It is a tough one, but I think it is the reverse side of the coin to the one you are presenting.

MEMBER: And also to make a decision as to a significant number.

DR. PALOMARES: Are we all together on the question? My feeling -- and I think I represent something of a unified feeling among colleagues of mine -- is that assimilation has been by default the key methodology of dealing with subgroups within our population. Some of us may feel that assimilation, in the way it is accomplished, may not be the best way, but we have not really developed a systematic way of communicating to other people the value of another way of developing into the "good American" or the "American who really counts."

I would like to propose that we begin to review the business of assimilation for succeeding in the culture, which is what the person eventually has to do. The Sioux Indian living in the Badlands of South Dakota sooner or later has to deal with society around him and that is one problem.

We are beginning to find out, more and more, through educational and psychologically relevant research, that the best way to assimilate a person may not necessarily be to ignore his particular culture; that the way to make me a good citizen of this country, a productive taxpaying individual, et cetera, et cetera, et cetera, may not be to ignore the fact that I am a Chicano. I pick crops, I speak Spanish, I have another culture, and have a whole world of my own.

By default, we assimilated such kids -- Italians and all of us came from other origins, excepting the Indians, of course -- in a fashion that we are beginning to question seriously because of educational and psychologically relevant research.

The idea was that you more or less ignored the background at best, or you actually chose systematically to destroy it or to punish him for it. If he speaks Italian, the parents say, "Don't speak Italian; that is the old country. We are Americans, now, and you speak American."

The Mexican-American parent doesn't teach the kids Spanish and doesn't teach them about the culture or about his background. In America, this system of assimilation worked well enough for many years, except that there are some populations that have doggedly resisted. They were here before the Italians or the Polish people or all of the people that led to you -- whoever they may be! It always sneaks out! And these populations are the Indian populations, the Mexican-Americans, and now the Puerto Ricans are involved. The alternative way of looking at it that we consider more educationally relevant is this. The whole bit is that the best way to make me a full and productive citizen is, first of all, to start where I am and make me proud of who I am, my language, my background, my parents.

If anybody destroys my language or downgrades my background or ignores me as a five- or six-year old, I see it as a rejection of me, as a person, and I do not distinguish between the language and who I am. The best way to build a good American is to build a good person, whatever his ethnic or cultural background may be.

The shortest distance to teaching a kid good English is to teach him well whatever language he speaks at home, and then he will learn English faster. To the degree that you downgrade his own language, you will be unable to help him come up in his education, generally.

MEMBER: When do you start that process?

DR. PALOMARES: That really threw me! Schools should, psychologically and educationally speaking -- not politically speaking because that is another question -- start as soon as the child enters school. I don't want to get involved in the theory of bilingual education right now, because that goes in a direction all its own. We haven't used that because we felt that it would take too much time, too much engineering, and too much money. As a matter of fact, we have been paying through the nose for these people in welfare, unemployment, jails, riots, far more than we would have paid if we had started the programs from the beginning. That is my feeling about the first part, but there was a second part.

MEMBER: When is "significant" significant?

DR. PALOMARES: As far as I am concerned, the adage we, as educators, have used for years (and which has not been implemented but should be) is: "start where the child is, whoever that child is." If he happens to be Indian and speaks a mixture of his language and yours -- mostly yours -- but still doesn't understand, then you as an educator owe it to that kid to go where he is.

I know what you are saying. We have to be realistic. Of course, we are not there yet but we are actively moving in that direction. But let us not kid ourselves by saying that we are offering equal education, because we are not. We simply are not. I know this seems impossible, because many of you have not thought about the business of individualized instruction, the business of starting where the child is and meeting his needs, treating him as a psychologically and educationally whole entity and dealing with him in this way instead of on a political basis.

"This is America and these damn kids should speak English. Kid, you speak English." You are making him pay for his parents. So, when is it significant? When one individual child has a need and it is different.

MEMBER: Give an estimated number. I think we are getting extremely emotional.

DR. PALOMARES: Right. For example, this is why I am opening up to questions.

MEMBER: According to you we would have to have as many different systems in particular groups as there are children. They can belong to many subgroups. I don't want my subgroup left out!

DR. PALOMARES: What I am saying is that my ideal of an education, as a reality, is that we should deal with each child individually, and we are not doing it. Take a look at what has happened.

MEMBER: It is just not practical, the way you are describing it.

DR. PALOMARES: Wait a second. It is not practical economically because of our value system of economics but not because it cannot be done by educators who really care. We have to make compromises, but let us not sell out completely.

The reason it is impractical to deal with each child as an individual within the environment and start from there is not because of the lack of money but because our values indicate we are not going to invest that much.

MEMBER: How about parameters like religion?

DR. PALOMARES: You know, the feeling I get from you right now is that you are baiting me.

MEMBER: No, I am trying to get a true understanding of your thesis, which is fascinating.

DR. PALOMARES: It is more than fascinating. It is going to be at your doorsteps in two months.

MEMBER: It may be on our doorsteps but we still have to have the wherewithal to carry it through.

DR. PALOMARES: You are right. I don't have the answer for each individual. Yes, religion is a part. Home, weather conditions, all are a part. All I am saying is that we have 30 percent, 20 percent of the kids in our schools who don't speak English well. I have gone to school after school that has had 30 percent of kids who have problems in speaking English.

Look at the total budgets of schools. Not one red cent is being spent on individualized problems. More time is being spent on the two percent than on all Mexican-Americans.

MEMBER: I will buy 20 percent, but that is a lot different from one out of a thousand.

DR. PALOMARES: I give up, you win. Next question.

MEMBER: You asked me at what point does it become legally significant. If you have one child you may have a case. In terms of litigation, where would you set the figures?

DR. PALOMARES: We are involved in the process, right now, of coming up with a minimum figure, and we will do so. It won't be a moral decision, but it will be a legally binding decision. Morally, I think that the reason we are having to do this legal business is because we didn't do the soul searching that we are now being forced to do.

I wish somebody had put educators in this position a long time ago so that we would have policed ourselves and the government wouldn't have had to step in. All of us knew those tests for mentally retarded kids were wrong. Every psychologist knew something was wrong with the test but no one had the guts to get a better one or do something about it. And now people are saying that the tests are no good and they

ought to be thrown out. This would be taking away a very valuable instrument from us. As I say, it is going to be a legislative, relevant decision more than an educational decision. What can I say? That is the way it is going.

MEMBER: Isn't it surprising that if the minority group happens to be the group to which we usually assimilate even if it is five percent, the program is aimed toward the minority group?

DR. PALOMARES: Did you hear that? Sometimes the minority group is five percent Anglo and the whole educational system is geared to them. If it is 95 percent Puerto Rican or 95 percent Mexican-American or 95 percent black, the program is still geared to the five percent.

I know what you are saying. We don't have programs for that other 95 percent. My only question is: when are we going to start making decisions that will move us in the direction of developing programs for the other 95 percent? The reason I am involved in legislation right now in a Civil Rights action is to get us to start doing that, because we haven't been doing it.

I am not paying attention to the hands. There are one, two, three, four ahead of you. I will take the lady in red.

MEMBER: I think you are settling for far too little, frankly. I think, in a district that looks at each child, the children are homogeneous to the extent that the district puts a value on the individual child and he will notice that these other groups really have people, not groups. Money isn't the issue, it is value.

DR. PALOMARES: I agree. She said we were settling for far too little. Every child should receive the same type of attention. It isn't the issue of money, it is the issue of value and the decision. I will tell you the way I view my situation now.

We are using, in the Mexican-American subgroup, the issue of unequal educational opportunities, because there are no different programs for them. We are a battering ram for your very own children and your very own lives and professions. Perhaps we will get more money to do your job, but we have got to start doing something amongst ourselves and not wait until the kids march down the streets and parents throw out the tests. We have got to start doing something ourselves.

There is a hand over here.

MEMBER: I don't think you can successfully legislate against ignorance, and I think the big problem is that good teachers have always done such things. Poor teachers have not, and I think it goes back to the system. Maybe you heard, in the previous presentation, that the schools were not being successful in other things, and -- if you have two million school teachers -- I don't know, most of them do these things because they are following rules and they misapply any rule, any law, if you don't have people there, and you can't legislate people's thinking.

DR. PALOMARES: I will tell you this, before I started in legislation I was trying to get people to think. I have gotten people to think much better by legislation on action than I was doing before.

Starting from our final point first, I am not naive enough to think that legislation is going to change people overnight. All I am saying is that legislation has the potential for getting people's noses to the grindstone long enough to start the processes moving slowly.

The best way to get kids to learn arithmetic is to practice it 30 minutes a day. If you leave them on their own to learn, 20 percent will learn and 80 percent will not spend the time on it. The way of getting almost 100 percent of the educators to pay attention to the issues is to keep the threat behind their backs that if they don't at least they are going to be harrassed. Up to now we haven't had that. Number two, about teachers, about educators, about me, about all of us, let me make a statement. Good teachers have been doing it all the time; bad teachers have not. Boy, that is a very frightening thing to me, theoretically. I don't know if I am talking to individuals any more. I use that to make a statement, but I want to use your statement to go out and make a point. I may not be talking to you any more. That is the statement: that good teachers have always done it and bad teachers, therefore, forgot it. Man, that is one of the most destructive, undermining statements that can be made about a beautiful bunch of people who are trying to grow. Oh, yes, good teachers can help bad teachers get better, too, and all get better in the process of helping these subgroups learn.

Good teachers, good as they are, have had a lot of subconscious unawareness of things they are doing that they don't even know they are doing. Let me give you a preview of a study now operating to get more data on this: let me show you what some very good educators -- you and I -- who are here, are doing right now, and leave out the bad ones who didn't come. Let us talk about us. We are the good teachers. I just want to make a statement and you are going to give me the answer.

If I like somebody, I tend to stand closer to him or her, distance-wise. If I don't, I tend to stand further away. If I like somebody, I tend to touch him more. If I don't, I touch him less. If I like someone, I focus on him more -- and less if I don't like him. Have you got that principle? We have done a study on this and our findings are ridiculous. We don't have large enough numbers yet. What happened was so shocking that I would like to share it with you. I won't talk about you. I will talk about me.

What we did was to take a class of kids of different colors, and all we did was to rank them according to color. And then we measured the distance the teachers tended to stand when interacting with them, teaching them, and so on. We measured the distance they stood from them. You give me the findings. The darker the kid, the further or closer the distance, or did it remain the same?

MEMBER: Further.

DR. PALOMARES: You know what you are saying? This is, theoretically you are probably not saying it for yourself. I'll tell you this. The answer is yes, dramatically. The darker the kid the less times touched or the more times touched? The less. There were exceptions. Okay. Second, the number of times talked to most? Again the same relationship.

I said, "Oh boy, those Gringos" -- that is you. Those Gringos are sure terrible. I am glad I am not that. They took a video tape of me reacting and guess what the relationship was, there. There were weaknesses in the study, weaknesses galore. We did this study during the art period, when it should have been done on the playground.

The teachers would say to us, "I do that? I don't do that to those kids. I am a good teacher; I love all of them the same." We say to the teachers, "Go and view and measure it." The teacher would say, "I spend all of my time with that dark one. I help him a lot. I spend more time with him than with anyone else." And then she sees herself spending perhaps 30 seconds talking to him very intently and then going and standing beside the other one and staying beside the lighter one. Her remembrance was that she spent 15 minutes with the darker one and 30 seconds with the lighter one. It was just the opposite.

Good teachers out there, unknown to themselves, are perpetrating this because of color, because of culture, because of the things that are built into all of us. We are perpetrating this on these kids and we don't want to own them. Let me tell you, the dumb kids tend to be the darker ones.

My only answer is: if you have darkness and dumbness what do you think you get? Let me make a final point. It goes something like this. Don't put it out that there are bad teachers who are doing worse. We good teachers, because of our inability to deal with ethnicity and race as a definite psychological and educational variable, have been perpetrating on other cultures astrocities and artificial educational values. It is time we started taking a critical look at ourselves and what we do to kids.

When that kid walks out, "Hey kid, what is your problem?"

He says the teacher doesn't like him. Then you ask the teacher, "You don't like him?"

She says, "I love him." Maybe what that kid should say is, "Miss, Mr., whatever your name is, I know you love me, but relatively speaking you stand further from me, you touch me less, you focus on me less and give me, all in all, less attention and love than you give the other children and it is the relativity of that that is killing me. It would not matter if you treated us all the same, but you don't and you don't know it."

By the way, that is us, not them. We are the good teachers. I know what every one of you is saying. Not I, I am the extra special one. Watch yourself on video tape.

MEMBER: This will change the subject, but has our government studied anything the Canadians have done? As I understand it, they entered with a different concept. Have we done any systematic study?

DR. PALOMARES: Especially with the French in Quebec, and so on. I didn't have to say that. No, not in terms of the work we are doing. This is something we are suggesting -- that we take a look at what systematically happens. They made this decision six years ago and are now deciding to convert back to a bicultural approach.

MEMBER: I don't know -- can you hear me? I hope what I am saying isn't out of order, but when you mentioned Arizona, you may have been one of the kids involved a few years back. I worked in the schools in Phoenix, in the downtown district, where we had a great many Mexican-American children and a great many Negro children and the woman who really ran things with an iron hand -- she had been there three years and retired last year -- placed teachers and decided where they would be assigned, where they would teach, when another teacher would be added to reduce the size of the class, and so on. In other words, she made all the decisions. I remember, in one school, on half-day sessions the first grade got up to 48 per class. That same week, at the north end of the district, a class got up to 41; a teacher was hired that morning and the class was split.

I was talking about the fact that these are the kids who need it. These are the ones who don't get help at home. And she patted me on the arm and said, "I don't know why you worry about these children. They will never amount to anything anyway." If you think anything he has said is overdrawn, you need to get into a situation like that and see it happening and try to fight it. And find yourself not in favor politically, you see, because you stand up for that kind of problem. And in assigning teachers, sometimes teachers are transferred almost as a disciplinary measure.

DR. PALOMARES: The salt mines.

MEMBER: You have really understated the case, probably because you didn't know how bad it really was.

DR. PALOMARES: Thank you.

MEMBER: The state law at that time mandated that all instruction must be in English and to get a teacher with 90 to 95 kindergarten in two sessions in one day -- most of them not speaking a word of English --

DR. PALOMARES: And legally bound not to teach, even if she could.

MEMBER: I had brilliantly learned four or five Spanish words and here a little fellow in the first grade was trying to read something to do with "house." He didn't understand, and I said "casa" -- I thought that was the word. "Oh, yes." He understood.

The teacher said, "You are not allowed to use a Spanish word."

I said, "You can talk for five hours and he wouldn't understand, but say that one word and he knows what it means."

This is the kind of senseless stuff that has been done by people and by the law.

DR. PALOMARES: That teacher, that person that passed that legislation may have been -- you know there are people out there -- some of you out there right now are making me feel bad. This is America and we speak English. You are saying why all of this cultural and language stuff? Looking at it from the economic angle alone, it would be much cheaper to deal with that when they are in kindergarten than when they are on the street, in jail, or on welfare.

MEMBER: What plans do you have for implementation of individualization with the CR and D class? How will this respond to your minority group?

DR. PALOMARES: Specifically, what we are trying to do is to go back and remediate some of the educational errors that occurred -- the atrocities in the misplacement of this type of child. Let us talk about the Mexican-American.

What is being done? Now we talk about accountability -- forcing educators to take the advice of other people, but this isn't for sure, we are just discussing it -- probably in a couple of months it will be solidified and sent to Richardson. The old system of screening where the psychologist reported the findings to the board is one of the reasons why the psychologist and the system have often misplaced kids.

What we need is a broader consideration of the problem. Let us talk about me. The way the thinking is going is that it is better to include a committee that meets and has in that committee community people and that the teacher, when she refers the child, is forced to go and see the child outside the school environment, at home, because one of the things that happens is that the child will often evidence all of the mentally retarded behavior patterns when associated with Anglo people.

So we say that part of it will be that the teacher referring the child to this committee will have to visit the home and get a report from there as to what is going on, because we find sometimes, that teachers will walk in, and there the kid will be buying all of the groceries at seven years of age. He does the babysitting, can add and substract, and is perfectly adjusted. We think this would cut out a lot of the referring. I know they are already supposed to be doing this. We hope, with the composition of the committee, the way things will be done will force them to do more, and not just maybe. Many of the schools have rules by which the teacher is not allowed to go to the home. Th is going to have to be changed.

The other thing is that when children are referred to the committee, this will be done on the advice of people from several different sources, before the kid is tested. Then, when the psychologist has done the testing, this is brought to the attention of the committee. By the way, I am a school psychologist!

When it comes to this type of child, he will have to mention to the committee the type of processes and assessments, the devices that he is going to be employing, and after he employs these he will be held accountable for using those devices. We are also going to try to back up the psychologist, and if he needs extra time the school cannot land on him. He should be free to do his job. We are hoping to be administered financially through Federal funding because we know the realities of the problem. This is one way that we hope to start. Beyond that we go into instrumentation. That is the way we are talking. That is another direction. That is about where we stand. We still have a lot of work to do. Does that speak to your problem?

MEMBER: What plans do you have for the educational program within the classroom?

DR. PALOMARES: We have none. Legally, it is difficult to step into that area because every local school has the right to control that, and it is difficult to mandate. We are not now in the process of legislating an educational program in terms of mental retardation. We hope to make recommendations about Mexican-American and other ethnic groups, but with the mentally retarded we are not there yet.

MEMBER: What recommendations will you be making?

DR. PALOMARES: In my thinking -- maybe you can start thinking about this -- we are thinking of saying the school should have a program. At least, it is moving towards the inclusion of programs to meet special needs like those of the Mexican-American with second-language programs and other programs of this type, and, if the school shows the proper direction in that area, that is all we ask. There are problems in that, and I will be the first to own it. As the man says, you can't legislate good education.

MEMBER: Getting back to the psychologist: when he has to say what kind of test he is using, do you have any suggestions there? Do you know of any that could be used?

DR. PALOMARES: Did you notice what I said? I said, what kind of assessment processes, what kind of methodology, not necessarily tests he is using with which to build, because there is this type of child who does need special placement? There is a battery of systems for taking a critical look at this type of child, but there is no specific test.

I don't know --- these Spanish versions of the Stanford, Binet, and so forth --- boy! what researchers have found out, and what I found out, was that the items still carry cultural bias. And the kid who is both linguistically and culturally handicapped usually scores lower on this type of instrument than on others. This is why I don't recommend any specific test. We are talking about other processes, including pretty complex matters, but I think it does a better job than what we are doing now.

MEMBER: I admire your enthusiasm, but I really question the validity of cultures sitting side by side and greater overall culture relating. In Quebec, French teachers are teaching French students to hate Anglo-Saxons while at the same time taking the Italian subculture and trying to assimilate it into the French culture.

I would like to know the basis for this optimism. I can see we need improvement but what evidence do you have that the alternatives for which you are legislating will work in the long run? Is that controversial enough?

DR. PALOMARES: I feel like a little boat sailing along with full sails and now the big one is crumpled a little. Why the optimism? A while ago a lady said that maybe you don't know the realities. I do know the realities. I was held three years in the second grade, graduated when I was 20 years old.

Education has been a struggle for me. I picked crops throughout the Southwest. Nobody knows more than I do about these difficulties because of my migrant background, because of being unable to speak Eng'ish -- I can honestly tell you the optimism I show comes from the feeling I have

about people that, if you start where they are, they will go further than if you get hung up on history.

I don't know why I am optimistic. But I believe this is the direction to take. I believe this direction has more promise than the position we were in before. I don't have any other. I don't think it is going to solve problems overnight. Man, I can see a school superintendent in Texas handpicking his board and doing everything he wants!

I can see the Mexican-Americans trying to assimilate forcibly the Mexicans that come from Mexico and are not Chicanos. I can see blacks forcing Chicanos to become black because we all have the same problem. When Martin Luther King ended his talk by saying the problem of the minority was the problem of the majority, he really meant majority. He meant black, white, and brown. As I said, I value whiteness over darkness, too. We all have the same problem. Why is this approach better than any other? I feel it is better because it gives the child a chance to develop a sense of faith in himself, within the world he lives in. Nobody is telling him that what he is is bad.

I do know, from sociology and psychology, that if a person is allowed and helped to believe in himself as a worthwhile person he has a better chance to succeed. To the degree that you destroy what he is through culture or language, you move him away.

Let's try something different. Let us try something that at least has more educational and psychological relevance. The other is political. I think we assimilated people that way not because we were dealing with kids but because of the political beliefs of adults. I think this new approach is more educationally and psychologically sound and relevant and thus adds a new dimension and a better chance of success. I am overly optimistic perhaps, but that is the way I am.

I must not forget the women because they are a subgroup. While she is coming up here -- she is shy!

MEMBER: Don't you ever call me shy! (I wouldn't have come up to the microphone except he said that.) You fought your way through. You had a lot of things going against you for a long time. You dug in the fields and there must have been a lot of people who knocked you down. In your experiences there must be something that made you fight your way through. What can we do for the children to help them fight their way through the way you fought your way through?

DR. PALOMARES: Can I save that? I will take this one and then end on that.

MEMBER: One of the problems, it seems to me, in attempting to answer most of the questions is that you try to use an educational frame of reference -- yet all the problems you have introduced are essentially social problems. Obviously, we as educators tend to forget there is a world out there.

I don't know whether one can indeed answer whether we should treat ethnic groups in one way or another, educationally, unless one answers in terms of the kind of world you are talking about. I would suggest that in these types of meeting and these types of discussion we also address ourselves to the fact that education is tied inescapably to political and social forces. You indeed indicated this in much of what you said and I appreciated much of what I heard. However, I would like to indicate that educators -- and I am one of them -- have kept themselves too far removed from the things we considered outside of our field and in a sense have abrogated our responsibilities as citizens as well as educators.

There is a political election coming up on Tuesday. I would suggest that, if proper political candidates were elected in terms of values, if we got out and voted in the primaries and perhaps stood up for election ourselves, we would be more likely to approach the kind of world that most of us, as good people as well as good educators, would like to see. And I suspect most of our students that are on university campuses and seem somewhat disillusioned and upset really are upset because of this particular phenomenon they see. I must say they, themselves, forced me into a little introspection over the past year to try to do a few things myself outside the educational system. I would suggest we are very closely tied to a broader socio-political system. No longer can we talk of rationalization of educational processes.

DR. PALOMARES: Yes. I am in agreement with what he said so I will go to the previous question, about helping children to fight their way through. A lot of people have asked me that question. And the answer that I think people want me to give and would like me to say is that, like Horatio Alger, no matter what happened, I pulled myself up by my bookstraps -- I made it on my own. fortunately for me, I have proof of a lot of bad things but I'm also forced to remember a lot of good things, and I would like to say that the reason I am where I am is because a lot of people gave a damn. They treated me as a person and made me feel proud of exactly who I was.

They, in fact, taught me what I was. I thought a Mexican-American was some dumb guy at the Alamo where thousands of guys my color, speaking my language, looking like me, got wiped out by John Wayne and -- what was his name? And I remembered that but I remember there were teaching people along the way who said to me, "You are the people you come from. You are part Indian, you are part Spanish, you are from Mexico and your parents picked crops, built railroads; when they couldn't have slaves from the South come to the Southwest they rented slaves from Mexico and you were a rented slave, but did a damn good job. Your parents speak a beautiful language; and, pretty soon, people will speak more of that language than of English -- you speak it well. You speak Spanish, but the way you pronounce your words now in Spanish will lead you someday to speak English perfectly." People, I thought I was terrible. I wanted to be something I wasn't. Some people, though, gave me pride in who I was and they went out of their way.

But, boys it has paid off, because I pay more taxes right now than most of you do! And if those people had not gone a little bit out of their way to help me to appreciate what I was and take pride in who I was, I don't think I would be here today. Somebody taught me to be a person and thus I did not make it by myself.

People worked hard to get the bootstraps for me to pull myself up by. One of the problems America has is that many of us think we did it alone. I remember one time a lawyer was in the group and we were talking about programs of this nature. He stood up and said, "Dr. Palomares, why is it you present special programs, special treatment? Pretty soon you will want everyone to have special treatment. Nobody helped me."

He was Anglo, blond, and he was successful. "I made it on my own and nobody helped me. I didn't have special programs." Then, I tried to say that the program was geared for him, but he wouldn't listen. He went on and said, "You people, you want handouts, you want help." The wind was out of my sails when a miracle happened. Guess who was the headstart teacher and sitting in that corner of the room?

MEMBER: His wife?

DR. PALOMARES: His momma.

CHAIRMAN HERRICK: Dr. Palomares, you have added a great deal to this conference. It has certainly been a privilege to hear you speak today.

(Whereupon the session was recessed at 11:40 a.m.)

- - -

FRIDAY LUNCHEON SESSION

October 30, 1970

- - -

The Friday luncheon session of the Educational Records Bureau conference convened in the Terrace Suite of the Hotel Roosevelt, New York, N.Y., October 30, 1970, at 12:20 p.m., with William W. Turnbull, Chairman, presiding.

Neurological Bases of Education

Jose M. R. Delgado, M.D.
Department of Psychiatry
Yale University School of Medicine

The evolution of civilization may be characterized by two main accomplishments: liberation from ecological elements; and domination of ecological forces reaching a peak with the technological development of electronics, computers and split atoms which have placed awesome mechanized power at the disposal of man. With our present knowledge and technical capabilities, the future of civilized societies does not depend completely on natural chance as in the past but may be determined to a considerable extent by individual choice, and we should remember that decision making is a method process which depends on brain activity. In spite of remarkable material advances, our psychic life and emotional reactions are little known. Solutions to present social problems proposed by sociologists, religious organizations, experimental institutes and even the United Nations have had only limited success. This is in part related to the fact that the usual frames of reference for these solutions have been politics, economics, history, metaphysics, sociology and psychology while the basic cerebral mechanisms related to man's ideas, emotions,

hostilities, desires and pleasures have been ignored. Individual reactions are determined by environmental factors acting through sensory inputs upon neurophysiological processes and manifested through motor outputs as behavior. All these intervening elements must be taken into consideration to understand and to educate the responses of individuals. To consider the problem only from outside the organism, as it usually is done, is as inadequate as if we ignored the environment and attempted to explain behavior solely in terms of neurologic activity.

The principle of education is always the same: to provide certain sensory inputs for the child, mainly auditory and optic, with the expectation that desirable patterns of behavior will be obtained. The teacher, however, is an outsider in the process of education. He faces the student audience and seeks to communicate knowledge without knowing what is actually going on inside the minds of his listeners. The professor provides the input of materials to be learned or "discovered" by the pupils, and he observes the results and assesses the intellectual output by grading the work performed. Unfortunately teachers know little about the mechanisms operating inside the brain, which is the essential link between inputs and outputs, or about the neurophysiology of attention, motivation, elaboration, information storage, recall, ideological associations, behavioral expression and other aspects of the phenomenon of learning.

The teacher has not only been unfamiliar with these mechanisms but has considered them out of reach. Pedagogic theories and experimentation in schools have attempted to provide scientific bases for the art of teaching, but results are often controversial and the establishment of controls has been especially difficult. Many of these problems derive partly from the limited help that pedagogy has received from psychology which until recently concentrated its research on the classical stimulus -- response relation, ignoring the study of intracerebral mechanisms. If we could explore the depths of the central nervous system while subjects were learning, thinking, or responding, perhaps we could detect the actual flow of electrically coded information within neuronic circuitry. These are precisely the possibilities which have now been introduced by the recent development of specialized bioelectronic methodology.

The technical breakthrough in the study of the brain in behaving animals came in the 1930's in Switzerland where Professor Hess implanted fine metallic wires within the brain substance of cats and demonstrated that movements of the body, sleep, rage, fear and other manifestations could be evoked by sending a few volts of electricity into specific cerebral areas. In these experiments, for the first time in history, typical psychological responses like offensive-defensive behavior were induced by direct application of electrical current to the central nervous system. Surprisingly enough, these findings, which were rewarded with the Nobel Prize, did not produce a significant impact on philosophical thinking and attracted only a limited interest among biologists until the middle 1950's when there was a sudden expansion of psychosurgery, psychopharmacology and physiological psychology. Many investigators realized the great importance of exploring the depths of the brain in awake subjects and started using intracerebral electrodes in animals and human patients. The idea of leaving wires inside the living brain may seem uncomfortable and dangerous, but actually hundreds of patients have walked around for days or even months with electrodes in their heads without any discomfort or ill effects. More recently other methods have been developed to apply chemical stimulations, to block neuronic structures, to record electrical activity, to collect neurohumors, to measure temperature, and to study several other phenomena of the behaving brain.

Among the many interesting results of these investigations, the demonstration that learning may be correlated with the functional activity of neurons is of special interest. It is well known that the brain spontaneously generates electrical waves which can be recorded by means of electrodes applied to the scalp or introduced into the nervous tissue, and investigators have identified specific electrical activity related to the reception of sensory stimulation, to the temporal association of two stimuli which represent the elemental basis of learning, and even to the exercise of mental activity such as solving a mathematical problem. With refined microelectrode techniques it has been possible to correlate the activity of single neurons with sensory reception. For example, in the occipital lobe of the cat, there are neurons which respond specifically to horizontal or vertical movements, to edges, patterns or colors.

It is generally accepted that not only the cortex but also the hippocampus and reticular formation are closely related with attention, recognition of meaning, learning and conditioning. The process of learning depends on the formation of new links between previously unconnected nerve cells by means of synaptic changes or perhaps by the establishment of new circuits. In addition, a state of increased motivation or attention is required to prepare the neurons for the functional -- and perhaps microanatomical -- changes which constitute the physiological correlations of learning. Experimental findings are still limited, and we know little about the anatomical location, chemical phenomena and electrical processes involved in the reception of information or in the storage and scanning of messages, but useful techniques for intracerebral studies are now available and are only waiting to be used by more investigators.

More is known about responses evoked by electrical stimulation of the brain, and there is evidence that most autonomic functions may be influenced by direct excitation of specific areas. This fact is important because of the well-known relations between psychic activities and vegetative functions such as respiration, heart rate, blood pressure, gastric secretion, intestinal motility and other autonomic manifestations. The diameter of the pupil, for example, can be precisely controlled like the diaphragm of a camera by adjusting the intensity of the electrical current applied to the hypothalamus, and the effect may be prolonged for as long as desired. One monkey equipped with a small transistorized stimulator attached to its collar remained free in its home cage with the right pupil permanently constricted throughout 21 days of hypothalamic stimulation, suggesting that the effect could be maintained indefinitely without fatigue. During this period, the pupil could still react to light but it was always smaller than the left one, demonstrating that brain stimulation evoked a functional bias without blocking the normal responses of the activated region. This experiment introduces the possibility of establishing a permanent bias within the brain by electricity or drug administration in order to modify undesirable emotional or functional states.

Many other motor responses have been induced in monkeys, cats, and other species by stimulation of determined cerebral structures (see Fig. 1 and 2) and the animals have been induced to turn the head, wiggle the ears, chew, eat, walk around, close the eyes, lie on the floor and perform a wide variety of movements which in general were well organized and often appeared directed by the animals' will for some useful purpose. For example, when licking was elicited in cats by excitation of the motor cortex, they looked actively for something to lick such as milk in a cup, the floor, their own fur or even the experimenter's hands. These stimulations were not uncomfortable, and on the contrary the cats seemed to enjoy the attention paid to them and usually rubbed against the observers and purred happily.

Behavior depends not only on the activation of some motor mechanisms but also on the inhibition of many other unrelated responses. To act is to choose one motor pattern from among the many available possibilities, and as we are well aware, inhibitions are continuously acting to suppress inappropriate or socially unacceptable activities. The education of children is largely based on inculcating patterns of behavior, teaching them what to do and what to suppress. It is therefore logical that the brain has powerful inhibitory areas which can be identified through experimentation. It has been shown that the normally voracious appetite for bananas disappeared in monkeys as soon as the head of the caudate nucleus was electrically stimulated. The animals closed their mouths and lost interest in the fruit, but were perfectly awake and alert. In similar experiments performed in monkey colonies, we have observed that during stimulation the animal actively rejected the banana and walked away. Various types of inhibition have been evoked in different situations including the loss of leadership in which a boss monkey was tamed and his hierarchial rank reduced following stimulation of the caudate nucleus. In some cases, a submissive monkey learned to press a lever which triggered radio stimulation of the dominant animal in the group, diminishing his aggressive behavior. The fact that one animal in the group is able to control the behavior of another by instrumental means has obvious social implications.

Penfield and other neurosurgeons have stimulated the exposed brain of many patients during surgical interventions, and more recently electrodes have been implanted in the brains of patients for diagnosis and treatment of illness such as epilepsy, intractable pain and involuntary movements, providing the opportunity to duplicate many effects obtained in animals, and especially to investigate changes in emotions and in the thinking process evoked by intracerebral stimulations. Among many other results, these studies have produced recollections of the past, sensations of fear and threat of unknown danger, increase of friendliness, feelings of pleasure and happiness accompanied by giggling, laughter and humorous comments, perception of words and phrases, sensations that the present had already been experienced in the past, blocking of thoughts and other effects. These facts demonstrate that the study of mental functions can be approached by well controlled and repeatable experimentation. Many patients have already been helped by this new methodology and far greater benefits should be expected in the near future.

The studies described indicate recent orientations of brain research, and in particular the possibility of studying the

mind experimentally and of exploring its activity during all kinds of behavior such as aggression and learning -- to be, at last, on the inside looking out. Several universities have recognized the potentials of neurobehavioral studies and have created facilities to foster the development of this field, but perhaps the urgency for finding new ways to understand and direct human behavior requires a more general effort; for it may well be that survival of the human race depends on a greater awareness and better education of our own intelligence and psychic values. The tremendous power derived from domination of nature should not be directed by men not yet in command of their own brain power. Perhaps we should remember the lesson of animal evolution: Diplodocus, megaterious and other gigantic reptiles were at the peak of animal size and strength, but their brains were disproportionally small, and these magnificent beasts could not survive in a rapidly changing environment. Our powerful and industrial civilization should be parallelled by a mental and emotional evolution which would lead us into a psycho-civilized age by finding intelligent solutions to problems such as hostility and fear.

It would be naive and biased to think that studying brain physiology would solve the ideological and political conflicts of mankind. My only contention is that, if we understand the basic mechanisms of mental functions, we shall be in a much better position to educate the mind intelligently and to search for new, practical solutions in order to avoid the present individual and social problems of mankind. We are not helpless: we can think, plan and act -- but we must make the choice. We must decide whether we prefer to accumulate a few more thousand megatons of destructive power and travel a few more million miles into the space beyond the earth, or whether we care to take time to know more about the inner space of the mind and to civilize our barbaric psyche.

The thesis and conclusions of this article may be summarized as follows:

1. We live in a civilized society geared to increase its own mechanization and material development while neglecting man's psychic evolution and personal happiness. The present educational system reflects and maintains this situation.

2. This imbalance has been determined by technical factors because we had methods for the exploration and use of natural powers while we lacked methods for the experimental investigation of our power source, the human brain.

3. In the last two decades, methodology has been developed for the investigation of the cerebral mechanisms related to mental functions in animals and man. Aggression, pleasure, fear, memory, learning, and other aspects of individual and social behavior have been evoked by direct stimulation of the brain.

4. The most urgent problem of our age is to improve our understanding of the human mind. We need to shift man's power, economic resources and education toward cerebral research, recognizing that the conquering of the mind is at least as important as the conquering of the moon.

5. The present curriculum should include the discipline of psychogenesis to provide information about the cerebral basis of behavior. The aims of a future psychocivilized society will be to increase social integration and individual differentiation. Personal happiness depends on environmental circumstances as well as on their interpretation by intracerebral mechanisms.

Bibliographic Note:

For further information about brain control, including technology, medial applications, educational implications, and philosophical discussion, the following book may be consulted: Physical Control of the Mind: Toward A Psychocivilized Society. by Jose M. R. Delgado, M.D., Vol. XLI, World Perspectives Series, R.N. Anshen (ed), Harper & Row, New York, N.Y., 280 pp., 1969.

Figure 1 Aggressive behavior may be induced in gibbons by radiostimulation of central grey and other specific areas of the brain.

Figure 2 Behavioral inhibition is determined in the gibbons by radiostimulation of the caudate nucleus.

FRIDAY AFTERNOON SESSIONS

October 30, 1970

ERB Co-Sponsored Sessions with
International Reading Association and
National Council on Measurement in Education

The Friday afternoon sessions of the 35th Annual Educational Conference were co-sponsored sessions with International Reading Association (IRA) and the National Council on Measurement in Education (NCME). Both sessions convened in the Grand Ballroom of the Hotel Roosevelt with the IRA session called to order by Chairman Ralph Staiger at 2:00 p.m. and the NCME session called to order by Chairman Elizabeth L. Hagen at 3:30 p.m.

CHAIRMAN RALPH STAIGER: I would like to call this meeting to order. I realize we are starting just a little late and perhaps we can move along briskly. Welcome to the co-sponsored meeting of the Educational Records Bureau and the International Reading Association.

It is perhaps fitting that the IRA will be involved in the ERB meeting because one of the parent organizations from whence the IRA stems was born right here at one of these meetings at the Roosevelt Hotel -- the National Association of Remedial Teachers.

And we are quite proud of our parentage in this meeting. You all have programs. I shall not attempt to read the program to you. You know that one of our speakers is Roger Farr, of the University of Indiana. He is the director of the Reading Clinic in the Institute for Child Study. He has been a teacher and has taught English, corrective and remedial classes. He has done extensive consulting and advisory work. He has been the editor of The Reading Research Quarterly and has done a great deal of work in the field of measurement in reading and we are particularly proud of one of his latest contributions, Reading Which Can Be Measured.

I hold the book before you tantalizingly. It is a remarkable compendium summary of research in the field of testing and reading and it contains a very useful index of the tests that are available, or were available when this volume was published early this year. This would be a very useful tool for anyone interested in testing. Dr. Farr will be joined on this program by Walter H. MacGinitie of Teachers College, Columbia University, who is also well known to you. He is professor of psychology and education at Teachers College, Columbia University, and has also taught at Long Beach, California. He has been a teacher in many different aspects, many different places. He is a research associate for the Lexington School for the Deaf in New York City and is the author of many articles and co-author of the Gates-MacGinitie Reading Tests.

His honorary awards include Honors in Mathematics, University of California; appointment as a MacMillan Fellow, Teachers College, and as a Life Member, California Congress of Parents and Teachers.

He has been a member of most of the important professional associations. I think it is not necessary but desirable to know that both of these talks this afternoon will be geared to the person who uses tests for instructional purposes. They have indicated to me that they will not be trying to speak to the test specialist, to the person who is not likely to use tests in the classroom.

Dr. Farr will speak first and I am delighted to present him to you.

DR. ROGER FARR: My major theme is that testing for decision making is what testing ought to be all about. If we don't delineate decision making situations and indicate what we want to know, we ought not to test. If that rule of thumb were applied in the United States today I suspect that a great number of testing programs would be cut in half or less. The problem with testing is that we are very confused as to what we want to test. If you were to ask a group of reading specialists to list the essential ingredients of reading ability, you would get quite different lists from each of these specialists. Until we can tell the test builders and publishers what it is we want to know, we are going to remain in a quandary and we shall also continue to be very critical of the tests when it is not the tests' fault.

Tests are essential not only to the reading program but also to the total school program. However, tests have been misused so widely that there has been an outcry against tests and testing in our schools. "Ban tests," is a slogan which might well be taken seriously unless present testing practices are changed.

I do not intend to focus my remarks on a critique of

present practices. Instead, I intend to suggest procedures for the effective use of tests in the instructional program. One of the reasons for the misuse of tests stems from confusion as to what tests actually are and what kinds of information can be expected from them.

Tests have often been called "measures of ability," "assessments of how much a student knows," "predictors of success," and the like. Such descriptions may or may not have something to do with testing, but they certainly do get at the essentials of what tests are.

Tests are a means of sampling a student's behavior under a given set of conditions (that is, the testing conditions). Each test produces a different sample of behavior under a different set of conditions. For purposes of my talk this afternoon, I am classifying tests as being either formal or informal.

Formal tests refer to standardized group or individual tests which follow standardized administration procedures and which usually provide norming data for comparative purposes. Informal tests are usually teacher constructed and range from informal reading inventories to work samples and observation forms.

Informal tests are very flexible to the instructional program. The teacher can devise them as needed to supply the information that the teacher decides would be helpful, given immediate instructional needs and, in addition, they can be used contiguously with instruction.

Thus, testing refers to the process of obtaining information about student behavior. There is a wide variety of forms in which that information can be obtained. Each has strengths and weaknesses, but the effective use of any of these means depends on the teacher knowing why she wants to test.

Thus, the first step in test selection is to define what information is needed and why it is needed. For instance, teachers often want to use a test to know: "Why isn't Johnny learning to read?" Is this question an adequate starting point from which to proceed to testing? My response is that it is not.

Teachers must know why they want to test in definite terms before they can test at all. They must ask more specific questions like: "Does Johnny have the vocabulary skills necessary to read Ivanhoe?" Rather than "Why can't he read Ivanhoe?" The essential question for teachers to ask is: "What do I want to know?"

Teachers usually have one of four reasons for wanting the information that reading tests provide. They want to determine a student's reading level. They want to assess his progress in reading development; they want to determine his reading subskills in order to form instructional plans to develop a more powerful reader; and they want to place him in the appropriate group for instruction. These, of course, are not the only reasons for testing. There are administrative purposes which have as their aim the evaluation of a school reading program. I do not intend to discuss these administrative reasons for testing; I am going to discuss the instructional uses of tests.

A word should be said, before proceeding to a discussion of the instructional uses of tests, about the kinds of caution which should be exercised in any use of tests. First, tests can only provide knowledge of the behaviors that they have been devised to sample. A test of vocabulary can provide information about vocabulary but not about reading speed, flexibility, comprehension, and the like. For a test to be useful, then, it must match the teacher's instructional objectives. In practice, this happens very seldom.

There are other misuses of tests -- having little to do with the tests themselves -- of which the teacher should be aware. Most of us tend to put great faith in anything "scientific" and are impressed by a series of scores which can be derived from some tests. We must be careful not to rely on numbers too heavily. When choosing a test, our own definitions of what we want to know are more important than whether the test affords an impressive set of scores. We don't always agree on our definition of what we want, to teach and measure, and I am not suggesting that we should

Another alarming misuse of tests comes from our tendency to treat tests as being predictive, rather than reflective, of a state of affairs. If a child performs poorly on a test of reading readiness, the task of the teacher is to develop that child's readiness skills in such a manner that the test is not predictive. All too often test scores are used to predict failure rather than to prevent it.

Finally, many tests, especially standardized tests, should be used cautiously because of the variety of ways in which reading is defined. Most standardized tests contain subskills, which vary from test to test.

Many times a test labeled vocabulary may not in fact be a vocabulary test but, instead, may be a test for other things like reading speed, word recognition, and so forth. Teachers must learn to rely on their own definitions of these skills rather than on the test definitions. Also, many tests which purport to measure the same skill are, in fact, measuring different things so that vocabulary measured by one test is different from vocabulary measured by another test. These problems should be considered when using tests to plan instruction. But the key to the effective use of tests is the teacher's knowledge of why she wants to test.

Teaching children to read is and should be a process involving continuous decision making, not only by teachers but by many different persons and agencies. State Departments of Education decide upon teachers certification requirements. Local boards, superintendents, principals, reading coordinators, and consultants work cooperatively to decide on the best use of available money, the need for special teachers, the suitability of particular programs, and so on. The classroom teacher is also vitally concerned with these questions.

Perhaps more importantly, however, she makes crucial decisions regarding the individual child. She determines which skills are to be taught to specific children and what are the functional reading levels and interests of each child. The list of instructional decisions made by classroom teachers is infinite.

The decision situation is not new to educators. However, most instructional decisions are made by forfeit; that is, by not recognizing that a decision can be made or by not being aware of possible alternatives. The usual forfeit "decision" involves continuation of a practice, whether or not it is the most appropriate procedure for the situation. Other decisions are made on the basis of limited or biased information; or they are made after consulting "expert" opinion, with little regard to the need and problems of the specific situation.

While the administrative application of measurement is appropriate and vital, it is within the classroom that the most important decisions regarding instruction are made. It is in this setting that the potential exists for helping children to become competent, interested readers, or (on the other hand) for handicapping them in skills development and "turning them off" from reading as a lifetime habit.

Reading programs which emphasize flexible grouping, individualized instruction, and continuous pupil progress are compatible with and depend upon the dynamic concept of measurement being proposed here. They are likely to involve choice or decision points before, during, and after an instructional sequence, with each point emphasizing a different decision. At first glance, this decision process would seem little different from what is done many times each day in thousands of schools throughout the country. However, a careful examination of the usual classroom practices reveals that measurement devices are not used as an aid to making decisions.

One of the major impediments to this development is that much of the educational environment mitigates against precise descriptions of instruction objectives and identification of procedural alternatives. Despite the restrictions of the educational milieu, the development of rational instructional decision making is the responsibility of the classroom teacher.

In order to develop useful measurement programs, teachers must state specific behavioral objectives; they must relate these objectives to classroom procedures; they must recognize alternatives to procedures and objectives; they must develop criteria for making decisions; and they must develop various methods for collecting the information needed for making decisions.

The development of measurement programs, will provide the pratitioner with data for the professional know-how to remove some of the malfunctions that tend to dissipate instructional energy. How can reading tests serve the instructional needs of teachers? How can they be used to provide useful diagnostic information, determine reading levels, and assess growth in reading power?

The most efficient procedure for determining instructional grouping or for comparing students in general reading development is to use a group standardized reading test. The selection of the appropriate test should be made by comparing instructional objectives with test objectives and by selecting a test which has the broadest possible coverage. In using the test results, no attempt should be made to use subtest scores for diagnostic purposes.

Care should also be taken to ensure that the test is not too easy or too difficult for the more able or less able students. Finally, standardized tests are valid for comparing students only when the standardized administration procedures are carefully followed for all the students who are to be compared.

After the teacher has obtained some idea from the standardized tests about who the good, the average, and the poor readers are, the next step is to determine their functional reading levels. Standardized tests do not tell us functional reading levels. Functional reading levels can be determined by studying the relationship between a particular standardized reading test and an informal reading inventory. An informal reading inventory, developed by the classroom teacher and based on the classroom instructional materials, provides a very useful measure of each student's ability to read at increasingly difficult levels.

Most often overlooked in the use of informal reading inventories is their use as a daily, continuous part of reading instruction. By constantly being alert to each student's reading performance and applying the criteria for assessing informal reading performance, the teacher can adjust the instructional material to ensure continued student success.

After determining appropriate reading levels for students, the teacher's next concern relates to the diagnosis of reading skills development. The validity of the teacher's diagnosis of students' reading skills can be increased if he selects or develops measurement devices that assess those skills he considers most important for the students' reading skill development.

This would mean that the teacher needs to accummulate a variety of procedures and devices for gathering background for instructional decisions. In order to diagnose any behavior it is necessary to know what the basic component of that behavior are.

I would like to dwell for a moment on the lack of agreement as to what the basic components of reading are. Research has been far from conclusive in defining reading. Much of it has taken the form of factor analysis studies in which various kinds of tests -- for example, tests of reading ability as well as tests of language usage and general intelligence -- are administered to a group of students and the test results are then analyzed to determine basic components of the reading act.

Several researchers have attempted recently to define reading in psycholinguistic terms. Goodman has developed a theory of reading which accounts for the nature of language and the reader's psycholinguistic background. According to Goodman, reading is a form of information processing; it occurs when an individual selects and chooses from the information available to him in an attempt to decode graphic messages.

Thus, Goodman suggests that perhaps the reading process cannot be fragmented. Ryan and Semmel's 1969 review of recent psycholinguistic theories of reading substantiate Goodman's point of view. They conclude, and I quote:

> Research has demonstrated that the reader does not process print sequentially, but rather in a manner which reflects his use of language at every opportunity. Expectancies about syntax and semantics within contexts lead to hypotheses which can be confirmed (or disconfirmed) with only a small portion of the cues available in the text. Thus, not all the information needed by the reader is on the printed page -- nor are all the printed details needed by him.

If one were to extrapolate components of reading behavior from their psycholinguistic theories, they would probably include the reader's ability to use knowledge of written syntax, knowledge of words used in context, and knowledge of how to use phonological cues. I suspect that would be the demise of many of the present subtests of standardized reading tests.

Perhaps the psycholinguistic approach will provide a more viable definition of reading and lead to a more solid basis for test construction. It may well be that research may find, as several of the proponents of psycholinguistic theory have suggested, that attempts to define reading subskills on a group basis are fruitless. In that case, measurement in reading would have to be based on whether a reader has a strategy for decoding written messages and whether he understands reading as a communication process rather than whether he can simply decode written symbols, supply the meaning of words in isolation, or answer multiple choice questions based on a literal understanding of a selection.

Until research is carried out to develop tests which take into account the elements that psycholinguistic theories are finding central to reading ability, the teacher will still need to use present subtests of reading to evaluate reading ability, but this use of subtests must be tried cautiously.

Present reading tests can be helpful if the subtests are recognized merely as measuring the reader's different ways of interacting with printed messages and are taken together to represent a measure of the students' ability to utilize text material effectively. Subtests of present standardized reading tests are merely different ways of looking at students' achievement in using reading text effectively.

There are a number of key problems in using standardized reading tests. First, there is no consistent definition of the subskills constituting reading on present standardized tests. This leads to confusion concerning their discriminate validity. This confusion has filtered down to the classroom where teachers have been left in a quandary about how to proceed with instructions.

Although available diagnostic tests seem to be quite limited, teachers can still plan effective reading programs that meet the needs of their students. This has been the case and will continue to be the case as long as the practitioner is aware of the limitations of the various diagnostic tests and realizes that the tests probably at best represent an obstacle course for the students. The best diagnosis takes place when the teacher brings "enough sophistication to the test session to evaluate pupils' reading abilities and weaknesses as they succeed or fail" on these various test items.

Adequate criterion measures of reading achievement need to be delineated before diagnostic testing can be improved. Standardized tests usually compare a student's performance with that of a given norm group. What are needed are tests which compare a student's performance to a given criterion for adequate reading.

For example, at present, only vague notions exist about what "good" third grade reading is. Until such criteria, or, perhaps more importantly, criteria for determining reading levels adequate for "effective" citizenship for adults can be devised, the value of diagnostic tests will continue to be based more on the sophistication of the reading teacher than on the sophistication or the intrinsic value of the tests.

In the hands of a skilled teacher of reading, informal measurement procedures are the most valuable procedures for planning reading instruction. In using informal assessments of the students' reading skills in daily classroom situations, the teacher can evaluate the students' ability to apply their reading skills to various learning tasks. He can also learn about students' attitudes toward reading tasks, and their reading interests.

My major conclusion, from a review of the research literature on methods for the diagnosis of reading, is that much research is needed before definitive suggestions for classroom practice can be outlined. However, such a conclusion is scarcely helpful to the practitioner who is faced with the immediate problem of how to diagnose an individual student's reading ability. Research should demonstrate that no one method can solve his problem. Knowledge of the diagnosis of reading achievement is not so scant that the teacher need to be paralyzed. Given a variety of procedures, teachers can make a reasonably accurate assessment of students' skills, capabilities, and needs.

Student growth in reading skills is the single most important goal of the reading program. Probably the most valuable contribution that measurement devices can make to reading instruction is that of providing reliable, valid assessment of this growth. The need for such assessment cannot be overemphasized.

Most of the elements within the reading program -- the teaching procedures, the grouping practices, the curriculum structure, and even teacher capability -- are evaluated on the basis of student growth. While it is not proposed that student growth be the sole basis for evaluating the reading program, nonetheless it is the single most important variable to consider in assessing reading programs.

I would like to suggest five steps that I think might be helpful in improving the assessment of reading growth. These steps do not solve all of the problems of measuring change, but someone this morning seemed to indicate we ought to quit measuring change. We shall not do so; just as we shall not quit teaching; we shall continue, but we want to improve.

> (1) The practitioner should carefully define the reading skill or skills being taught and select a measuring instrument or several instruments that are operational definitions of these skills.
>
> (2) If test norms are used for comparisons, the test user should be sure that the norm group matches the group being tested in all important factors related to growth in reading. Developing a local norm is, for most purposes, the best procedure.
>
> (3) Measurement procedures should be used under conditions approximating those of the actual teaching situation as closely as possible. If instruction had been designed to produce a generalization of the skills, testing should be done under those conditions to which this skill will generalize.

(4) If students have been selected for a reading program on the basis of their performance on the lower extremes of test score distribution, some procedure such as the residual gain scores should be applied to remove regression effects.

(5) Evaluation of change scores should be interpreted cautiously. The irregular growth curves of individuals indicate that reading improvement is uneven and that measurement in reading always involves some error.

My talk has focused on the contribution which various procedures for measurement can make to the teaching of reading. Much of the research concerning the measurement of reading casts considerable doubt on the validity and reliability of all testing instruments in general and group standardized tests in particular.

This is not to say that measuring devices have no value in reading instruction. On the contrary, tests can make a valuable contribution to classroom practice if they are used with caution and if the test user is thoroughly aware of their limitations; the test consumer should know why he wants to test and what he wants to test. In addition, the objectives of the test and the objectives of the instructional program should be closely related.

I realize I have not provided detailed procedures for using tests in the school program. What I hope I have done is to provide guidelines. I hope I have indicated to you that you ought to take a careful look at your program and say, "Where do I make the decisions with this program and which decisions are so important that I ought to get information for making the decisions and then select a test to help me get the decision?"

If the guidelines seem sparse, it is because the state of knowledge in the field of testing and evaluation in reading is limited. In fact, present measurement practices and instruments often are not as helpful as they could be in teaching reading. This is not the fault of just test consumers or test producers alone. Test users have been naive about the value of tests in the classroom. This has led to gross misuse of tests and situations where important stated objectives of reading programs have consistently been unevaluated. More often that not, group standardized reading tests fail to provide teachers with information about students' instructional reading levels, basic reading skills development, or attitudes toward reading.

Most reputable test publishers do not claim that tests can supply such knowledge, but they do imply that they provide diagnostic information by including reading subtest profiles and grade level norms. Some test publishers are convinced that teachers believe "grade norms" means something in terms of students' reading performances.

Despite the fact that redundancy may reduce my effectiveness by "turning you off" to my suggestion, I would like to conclude my talk by stating that the single most important practice for improving the instructional use of tests is for teachers to identify the decision point in the instructional sequence. Test instruments, of a wide variety, can then be selected or developed to provide information for making those decisions. Despite this need, there is currently a dearth of guides for decision rules, a lack of appropriate measurement devices, and limited understanding of the nature of the reading process. However, these limitations will not halt reading instruction, and they should not prevent the development of measurement as a process for providing information for making decisions. It is quite possible that the plea for accountability will lead educators to accept inappropriate goals, procedures, and outcomes all based upon inappropriate measurement. The potential disaster of "commercial accountability" can be avoided only as teachers of reading address themselves to the problems of self-evaluation and self-improvement by providing evidence of classroom accountability.

(Whereupon the meeting was recessed at 3:40 o'clock p.m.)

"WHAT ARE WE TESTING IN READING?"

Walter H. MacGinite

Most standardized reading tests used in the United States are group tests, usually given by the classroom teacher or reading supervisor to an entire class at one time. They are also multiple-choice tests, designed for rapid and objective scoring, often by machine. Such tests are usually given for the purpose of estimating general achievement of both individual pupils and the class or school rather than for diagnosing specific reading difficulties of individuals. Diagnostic tests, usually individual tests and administered by a reading specialist, are also important and widely used, but I shall not discuss them further at this meeting.

Group-type reading achievement tests usually consist of at least two subtests -- a vocabulary subtest and a comprehension subtest. Other subtests are also often included -- for example, a test of reading speed. Or the vocabulary test may be subdivided into two different types of vocabulary test, or the comprehension subtest may be divided into two or more different types of comprehension test.

What is it that is being tested by these vocabulary and comprehension subtests and by the further breakdown of vocabulary or comprehension? The points I will make in answer to this question are not new, and they all seem fairly obvious, yet teachers and educational researchers, too, repeatedly form conclusions that ignore these points.

What is being measured by these vocabulary and comprehension subtests? The first point is that there is as much of a difference between different educational levels of the same subtest as there is between subtests with different names at the same level. The great changes that take place in arithmetic achievement tests from one grade to another are self-evident to most people. To score well on an arithmetic test for the sixth grade, a student must know a lot of things about decimals and fractions that have no bearing on performance on a test for the second grade. Most teachers and researchers are now also aware that what is measured by so-called intelligence tests changes considerably from the infant level to the intermediate grades. In contrast, the rather large change in the content of reading tests from the first to the later grades is frequently not taken into account. Although most people readily see, or already recognize, the different requirements posed by reading tests at different grade levels, they seem seldom to consider these differences when interpreting research findings or a child's educational status.

I will first try to characterize briefly the changes in reading tests that occur over the first few grades, then offer two reasons for our relative lack of awareness of these changes, and then mention some implications of these changes.

Grade changes in reading tests are most obvious in the vocabulary subtest. The easiest items for the first grade usually use simple words well known to all children in speech. The distractors, or wrong answers from which the children may choose, may all look and sound quite different from the right answer and be quite unrelated in meaning. In slightly harder questions, the distractors will present possible perceptual confusions, so that if the right answer is house, distractors might be horse or mouse, or if the right answer is rose, distractors might be rope or hose. The vocabulary questions gradually are made more difficult by using words that are less likely to be known as sight words or words that include more difficult letter combinations.

Eventually, as the items get more difficult, the main difficulty for most children comes from uncertainty about the meaning of the words. The majority of the older children can puzzle out the pronunciation of most of the words whose meanings they know. They can even give a reasonable pronunciation to nonsense words. The test maker simply runs out of meaningful possibilities for making items more difficult by means of perceptual similarities alone. But we recognize that, for an older child, having a good reading vocabulary means more than just being able to pronounce words. The developing student learns new word meanings that a few years ago were not familiar to him in speech. Some of these new words may even now be unfamiliar to him in speech, but their meaning is recognized in print. Thus, a reading vocabulary test for older children is more concerned with whether the child understands a variety of words that he may find in written material.

This change occurs gradually in tests intended for increasingly more able readers. The title of the test remains the same ("reading vocabulary" or whatever the testmaker chooses to call it), but the ability that is tested appears to change quite radically. As represented by the harder items in a third-grade test, or by the majority of items on a fourth grade test, the reading vocabulary test has evolved into a test that is nearly indistinguishable from the vocabulary section of many group intelligence tests. Thus, the correlation between a reading vocabulary subtest at the fourth-grade level and a verbal I.Q. test is likely to be as high as the correlation between the reading vocabulary subtest and a reading comprehension subtest.

Grade changes in reading comprehension tests roughly parallel those described for reading vocabulary subtests, though they are perhaps less drastic and less obvious. In the primary grades the comprehension tests are more concerned with the straightforward interpretation of concrete statements and relationships, often those that are easily pictured. Sentences are simple, the number of items to be related is limited, and items to be related are not widely separated in the text. For example, the child may read a question like "Who is reading from a little book?" and answer by choosing a picture. In later grades, greater stress is laid on inferences, on understanding complex ideas and difficult sentences, and on applying background knowledge. Even for third graders,

items that are fairly easy for those who have mastered the mechanics of reading may involve simple inferences about matters that are not explicitly stated. Consider, for example, the following paragraph and question.

> Yesterday Ellen phoned to ask if we could come play with her. We ran right over to her building and into the lobby. The elevator was slow; it stopped at almost every floor. When we finally stepped out at the tenth floor, Ellen was waiting for us by her door at the end of the hall.
>
> Ellen lives in
>
> a trailer an apartment a farmhouse

It is not explicitly stated that Ellen lives in an apartment, but the capable young reader has no difficulty in inferring it.

Since these grade changes in reading tests are so obvious -- particularly in the case of the vocabulary subtest -- why aren't they more prominent in our thinking about the meaning of reading test scores? I believe there are at least two reasons. We recognize the changes in the content of arithmetic tests partly because these changes reflect the formal introduction of specific topics in our teaching of arithmetic. We introduce long division or the addition of fractions as a specific topic of instruction. We don't expect the students to know much about these operations before they are formally taught, and, after they are taught, we expect to see them featured in arithmetic achievement tests. Except for the so-called decoding stage of reading instruction, we don't have such clear-cut ideas about separate topics in reading instruction. This situation is natural enough, for beyond the decoding stage, advancement in reading depends so much on the child's developing language abilities that interact with most all other instruction and skills, such as locating the main idea or understanding poetry, but we are relatively uncertain about how to teach such skills; they often seem to develop without specific instruction, and they are highly intercorrelated, a point I shall return to later.

A second reason that we are relatively unconcerned about grade changes in the content of reading tests is that the same children who learn the decoding skills readily also typically continue to score well on later tests of richness of vocabulary or inference. There is considerable evidence of this stability of performance. Studies by Joseph Breen, for example, show correlations generally in the 70's between reading achievement at the end of grade one or grade two and reading achievement in the fourth or fifth grade. Now such stability could be taken as evidence that the tasks posed by reading tests really do not change very much from first to fifth grade. I have, after all, offered the high correlation between intermediate-grade reading vocabulary tests and verbal aptitude tests as evidence that they are testing about the same thing. The difference in the two cases is partly the evidence of one's eyes. The reading vocabulary sections of a reading test and of a paper and pencil verbal aptitude test look alike. They were prepared following similar principles, to test, in printed form, richness of vocabulary. On the other hand, reading vocabulary and comprehension items for the early grades are built on different principles from those for later grades, as I have described already, and the result is readily apparent in the items.

There are other considerations, also, to make one reject the high correlation between first and fifth grade reading scores as evidence that items designed to test decoding skills are actually testing the same ability as later items. One of these considerations is that some of the variance in scores at first and second-grade level is based on items like those for higher grades. The harder items on second grade tests, at least, are often constructed like those for higher grades. The norms on such tests, after all, extend into the intermediate grade level. Again, this situation results partly from the fact that reading achievement for many children is not so dependent as achievement in some other subjects on specific school instruction.

There is another consideration that argues against accepting the high correlation between beginning reading achievement and later reading achievement as evidence that the beginning and later items are measuring the same reading skills. This consideration is that first and second-grade reading achievement, scores correlate remarkably highly with all kinds of later academic achievement, including arithmetic, not just with later reading achievement. In Breen's studies, mentioned earlier, correlations between first or second grade reading scores and fourth or fifth grade arithmetic scores were also in the 70's, though somewhat lower than correlations with fourth and fifth grade reading scores. Correlations between first or second grade reading scores and composite scores on the Iowa Tests of Basic Skills in fourth or fifth grade were in the 80's. The first or second grade reading items are clearly not arithmetic items. They are simply measuring something that is strongly related to later achievement.

It is interesting to speculate on the factors that are behind this relationship. Probably several factors are involved. I have discussed some of the possibilities in another context, but it will be worth digressing briefly to consider some of them.

Why are early reading scores so highly related to later school achievement? Do teachers continue to favor children who are initially favored by them? Do scores on early reading tests influence teachers' expectations and lead to self-fulfilling prophesies? Do homes that provide support for early success in reading continue to provide good support and encouragement for other school achievement? Do children who have the capacity to learn to read easily also have good capacity for other learning? Do children who are adaptable and malleable enough in the school environment to participate well in beginning reading instruction also participate well and thus learn more from later instruction? Does the reading skill itself, and the knowledge gained through using it, contribute so much to school achievement in other subjects that growth in achievement is essentially determined by it. Probably all these things, in varying degrees, are true. You can undoubtedly add other reasons to the list. My own belief is that, of the possibilities mentioned, perhaps the most important is the continuing and reasonably consistent influence of the home environment. There are great variations in the degree to which the home provides a source of motivation and support, establishes habits of attention and cooperation, provides a background of useful skills and information, and, probably not least in importance, supplies actual instruction on school subjects.

In any case, for whatever reasons, reading ability at the end of first or second grade is highly related to later achievement in reading and other subjects. Put another way, a child who has not learned to read by the end of the second grade is in deep trouble in most school systems. The child who does not learn to read in first or second grade finds that he has been planted in a child's garden of reverses. There are exceptions, of course, but most such children are in for a long career of frustration and failure. That there is a strong correlation between early success in reading and later school achievement does not necessarily mean that preventing early reading failures would drastically reduce later school failures. The effects of a prevention program would depend on the reasons for the strong relationship between early reading achievement and later school achievement. On the other hand, we know that if nothing is done, those children who now do not learn to read in the first two years are very likely to be saddled with failure for the rest of their school careers. It is surely worth a try -- worth an all-out effort to try to see that every child who doesn't make good progress in early reading has every incentive and every opportunity to learn the skill. I am not suggesting that all children can achieve equally well, simply that the school should recognize what an extremely serious matter it is when a child doesn't learn to read in the first grade or two and that the school should do all that possibly can be done at that time rather than waiting until later.

So far, I have been illustrating the point that the nature of reading achievement tests changes markedly from the first grade to the intermediate grades. Let us now look briefly at the other side of the statement that introduced this point, namely that, at a given educational level, there is not much difference between reading subtests with different names. Correlations between the vocabulary subtest and the comprehension subtest generally approach the reliability of the individual subtest. There is still room for the two subtests to be measuring somewhat different achievements, but, for individual pupils, the difference between the vocabulary score and the comprehension score must generally be very large before we can put much faith in this difference actually reflecting a true difference in achievement in the two areas. The same statement applies with even greater force to attempted subdivisions of the vocabulary and comprehension tests. At the intermediate grade level and above, repeated studies of different types of formats of vocabulary testing emphasize that more or less the same achievement is being measured by the different types of vocabulary tests. There is, indeed, some difference, but the value of separate subtest scores for different types of vocabulary test at the intermediate grade level and above seems questionable at this time.

At the stage of beginning reading, however, there is probably room for more differentiation of the skills that are tested than has so far been incorporated into most tests. Any achievement test should, of course, be directly relevant to what is being taught in the school. At the present time, there is a considerable variation in the way beginning reading is taught. Some programs emphasize a mastering of the graphemephoneme correspondences in English, while other programs also stress, at an early stage, the need to recognize many of the more common and useful words that are irregularly spelled. Most vocabulary tests for beginning readers include a mix of items for measuring the outcomes of both of these emphases. The result is that a child who is well on the way to mastering the decoding aspect of reading (in that he can pronounce a great many regularly spelled

words), may not fully show this strength on many current standardized reading tests. It would seem appropriate, therefore, to provide separate tests or subsections appropriate to each of these two emphases in beginning reading instruction. This arrangement would be moving in the direction of criterion-referenced measurement, which will be discussed in the next section of this program. Let me emphasize in this context that a child who only knows how to pronounce regularly spelled words is still at the very beginning stages of reading achievement. I believe it is at these earliest stages of reading instruction that criterion-referenced measurement can be most meaningful and helpful in assessing reading achievement at the present time. At advanced stages of achievement, criteria will be much harder to specify, and if we follow our intuitions in setting them, we are likely to obscure rather than clarify the problem of the taxonomy of reading ability. Some criteria that will seem to make common sense will not help us understand what skills we need to teach. We need to continue to study this problem of the skills and abilities that compose reading achievement.

At the intermediate and high levels, separation of different types of comprehension is about as difficult as separating different aspects of vocabulary achievement. The work of Fred Davis, who will be speaking in the panel discussion to follow, has been clarifying the nature of this problem and indicating some of the potentials that exist. At the present time, the most promising distinction, exclusive of vocabulary, would appear to be that between understanding facts explicitly stated in the reading passage, and making inferences from what is stated. Even this distinction is not an easy one, and we should require a clear demonstration that two subtests are measuring this distinction before we pay attention to comprehension subtest scores that claim to represent different aspects of comprehension ability.

Let me now illustrate the significance of the changes in the nature of reading tests from first grade to the later grades by giving two examples of how these changes might influence our understanding of research findings or test results. We have all been concerned with the gap in reading achievement between disadvantaged slum youngsters and their middleclass peers -- a gap that appears to increase the longer the youngsters are in school. One of several possible reasons for this increasing gap is related to the changes in the nature of the reading achievement test. It is quite possible that differences in home background are more influential in determining the score on the conceptual tasks of the later reading achievement tests than on the more perceptual tasks of the earlier tests. Surely, it is precisely in the area of the richness of the child's standard English vocabulary that we would expect home background to have one of its greatest influences. This factor, of course, does not rule out others, such as differences in the quality of teaching or the cumulative effects of motivational differences.

Looking now at another aspect of these changes in reading tests that result in their becoming increasingly more of a conceptual task. It was noted earlier that the reading vocabulary subtest ended up with the intermediate grades being essentially like the vocabulary section of a group intelligence test. Some cities have recently abandoned, or in fact, banned, the use of so-called intelligence tests in the school system, on the grounds that they lead to discrimination against pupils whose backgrounds have not equipped them well for traditional school studies. Should not the reading-vocabulary test then be banned as well? When one seeks to answer this question, it becomes evident that the potential harm from the intelligence test lay in its title and in the surplus meaning given to the scores, not in the information it actually provided. It provided information about the student's current ability to learn academic subjects through reading, or listening, to expositions of academic material in standard English. The reading-vocabulary test provides that kind of information, too. In fact, at the end of first or second grade, a combined reading vocabulary and reading comprehension test is likely to predict later school achievement more accurately than an I.Q. test will. But look at the difference in attitudes toward these two test scores. We ban one, but we give increasing attention to the other as an index of what the school has been able to accomplish. Yes, there is the difference. The reading-vocabulary test is looked on as a measure of the school's accomplishment or the school's failure, whereas the vocabulary section of an intelligence test yields a score that is someone else's responsibility. One way of indexing the difference in attitude toward the two types of tests is to note the difference in the temptation to coach students on the answers to the two. Coaching, and other fraudulent ways of making sure that the reading test scores of a class or school look good, has become a serious problem in some school systems. Coaching is ordinarily not a problem for I.Q. tests given by the school. If a slum school were to claim that the average I.Q. of its pupils was 100, it would produce a real crisis for the teachers. We would somehow expect the teachers to produce a level of pupil reading achievement that was up to the national average. On the other hand, if the average I.Q. in the school was 85, we tend to expect less in the way of achievement. The low I.Q. score is taken as an indication that the children will have difficulty in learning. It can even serve as an excuse.

The teacher may not realize that the reading vocabulary test is very like a section of the I.Q. test. But the teacher does know that a child who scores low on the reading test will have difficulty in learning at school, just as she knows that the child who scores low on an intelligence test is likely to have difficulty learning in school. The teacher will probably assume, however, that the difficulties have different sources and different remedies. She believes that the remedy for the low reading test score and for the difficulties that it indexes is to teach the child to read. The teacher is likely to see a low score on a reading test as meaning that she must teach the child to read. She is likely to see a low score on an I.Q. test as meaning that she can't teach the child to read.

My purpose in raising these questions about the similarity between reading-vocabulary and I.Q. vocabulary tests and about the difference in reaction to them is not to get the reading tests banned too. The reading part of a reading test is the comprehension subtest, and surely we do want to know how well children are learning to read. Rather, I wish to point out that similar experiences and similar background factors influence the scores on the reading-vocabulary test and on the vocabulary section of the I.Q. test.

In the past, we have tended to think of the intelligence test score as reflecting the child's past and as indicating the extent to which he will be a problem for the school in the future. We have thought of the reading test score as reflecting the school's success in the past and as indicating the extent to which the child will have trouble in the future. We will face more intelligently the tasks of teaching reading and will face with even greater determination the whole job of education when we understand the functions and problems of measurement well enough to realize that both scores reflect the child's past and the school's past success, and that both scores suggest future needs and opportunities for both the child and the school.

- - -

FRIDAY AFTERNOON SESSION

Session Two

October 30, 1970

- - -

ERB Co-Sponsored Sessions with
International Reading Association and
National Council on Measurement in Education

The Friday afternoon sessions of the 35th Annual Educational Conference were co-sponsored sessions with International Reading Association (IRA) and the National Council on Measurement in Education (NCME). Both sessions convened in the Grand Ballroom of the Hotel Roosevelt with the IRA session called to order by Chairman Ralph Staiger at 2:00 p.m. and the NCME session called to order by Chairman Elizabeth L. Hagen at 3:30 p.m.

"SOME LIMITATIONS OF CRITERION-REFERENCED MEASUREMENT"

Robert L. Ebel
Michigan State University

Every mental test is intended to indicate how much of some particular characteristic an individual can demonstrate. To determine and express "how much," one needs a quantitative scale. Even those tests used primarily for categorical pass-fail decisions almost always involve a quantitative scale on which a critical "passing" score has been defined. Because the human characteristics that mental tests seek to measure are often complex and hard to define, appropriate quantitative scales are not easy to establish. Some of the most difficult problems of mental measurements arise in the process of getting a useful scale.

The essential difference between norm-referenced and criterion-referenced measurements is in the quantitative scales used to express how much the individual can do. In norm-referenced measurement the scale is usually anchored in the middle on some average level of performance for a particular group of individuals. The units on the scale are usually a functional of the distribution of performances above and below the average level. In criterion-referenced measurement the scale is usually anchored at the extremities, a score at the top of the scale indicating complete or perfect mastery of some defined abilities, one at the bottom indicating complete absence of those abilities. The scale units consist of subdivisions of this total scale range.

It is interesting to note that the percent grades which were used almost universally in schools and colleges in this country up to about 1920 represent one type of criterion-referenced measurement. True, the extremities of the scales used for percent grades in most courses were very loosely anchored in very poorly defined specifications of what would constitute perfect mastery. But this lack was more a consequence of the great difficulty in developing such definitions than of failure to appreciate their importance. Little has happened to the subject matter of education since 1920 that would make the task of defining complete mastery any easier. If anything, as the scope of our educational content and objectives has broadened, the task has probably become more difficult.

Thus the replacement of norm-referenced measures by criterion-referenced measures in education is not likely to be easy. If it were to happen in the next decade, as some seem to advocate, educational measurement would have come full circle. Those who accept the half-truth that there is nothing new under the sun would have another example to cite. More importantly, the difficulties and limitations of criterion-referenced measures, which half a century ago led to their virtual abandonment, would once again become apparent and would in all probability start the pendulum swinging back toward norm-referenced measurements.

This is not to say or to imply that there is no value in criterion-referenced measurements, or no possibility of using them effectively. They have a kind of meaning, a very useful kind that norm-referenced measurements lack. In some instances good criterion-referenced measures can be obtained.[1] But it is to say that the idea of criterion-referenced measurement is not new, that recent emphasis on norm-referenced measurements has not been misplaced, and that good criterion-referenced measures may be practically unobtainable in many important areas of educational achievement.

Criterion-referenced measures of educational achievement, when valid ones can be obtained, tell us in meaningful terms what a man knows or can do. They do not tell us how good or how poor his level of knowledge or ability may be. Excellence or deficiency are necessarily relative concepts. They can not be defined in absolute terms. The four-minute mile represents excellence in distance running, not in terms of any absolute standards for human speed, but because so few are able to run as fast as that for as long as that.

Now in many areas of education we do pursue excellence. In many areas we are concerned with deficiency. For these purposes we need norm-referenced measures. To say that such measures leave us in the dark about <u>what</u> the student is good at doing or poor at doing is seldom a reasonable approximation to the true situation. Usually our knowledge of typical test or course content gives us at least a rough idea of amount of knowledge or degree of ability.

One limitation of criterion-referenced measures, then, is that they do not tell us all, or even the most important part, of what we need to know about educational achievement. Another is, as we have already suggested, that good criterion-referenced measures are often difficult to obtain. They require a degree of detail in the specification of objectives or outcomes that is quite unrealistic to expect and impractical to use, except at the most elementary levels of education.

The argument that effective teaching begins with a specification of objectives seems logical enough. If we will settle for statements of general objectives, unencumbered with the details of what is to be taught, how it is to be taught, or what elements of knowledge or ability are to be tested, it is practically useful. But general objectives will not suffice as a basis for criterion-referenced tests. And the formulation of specific objectives which would suffice costs more in time and effort than they are worth in most cases. Further, if they are really used, they are more likely to suppress than to stimulate effective teaching.

The good teacher knows and is able to do thousands of things that he hopes to help his students to know and become able to do. Some of them are recorded in the readings he assigns or in the lecture notes he uses. Others are stored in his memory bank for ready recall when the occasion arises. Why should he labor to translate all these detailed elements of achievement into statements of objectives? If he were to do so, how could he actually keep such a detailed array of statements in mind while teaching? And if he were to manage such a tour de force, how formal, rigid and dull his teaching would become!

There is obvious logic in the argument that teachers need to think hard about their objectives in teaching. But when the argument is extended to call for specific statements of objectives, written before the teaching begins, it involves assumptions and implications that are open to question. One is that instructional efforts are guided more effectively by explicit statements of objectives than by implicit perceptions of those objectives. Another is that the effectiveness of a teacher's efforts depends more on the explicitness than on the quality of his objectives, or that explicitness means quality where objectives are concerned. The implication is that programmed teaching which has been carefully planned in detail is likely to be better than more flexible, opportunistic teaching.

Have you ever seen a statement of objectives for educational achievement (not just an outline of learning tasks to be performed) which did justice to all the instructor actually taught in the course and which therefore provided a solid foundation for criterion-referenced measurements of achievement in the course? If you have, did you not find that these objectives substantially duplicated the instructional materials used in the course?

Criterion-referenced measurement may be practical in those few areas of achievement which focus on cultivation of a high degree of skill in the exercise of a limited number of abilities. In areas where the emphasis is on knowledge and understanding the effective use of criterion-referenced measurements seems much less likely. For knowledge and understanding consist of a complex fabric which owes its strength and beauty to an infinity of tiny fibers of relationship. Knowledge does not come in discrete chunks that can be defined and identified separately.

Another difficulty in the way of establishing meaningful criteria of achievement is that to be generally meaningful they must not be idiosyncratic. They must not represent the interests, values and standards of just one teacher. This calls for committees, meetings and long struggles to reach at least a verbal consensus, which in some cases serves only to conceal the unresolved disagreements in perceptions, values and standards. These processes involve so much time and trouble that most criterion-referenced-type measurements are idiosyncratic. Is this not what was mainly responsible for the great disagreements Starch and Elliott[2] found in their classic studies of the grading of examination papers? To the extent that criteria of achievement are idiosyncratic they lack validity and useful meaning.

So a second limitation of criterion-referenced measurement is the difficulty of basing such measurement soundly on adequate criteria of achievement. The third and final limitation to be discussed here is less a limitation of the method of measurement itself than of one of the principal justifications that has been offered for its use. This justification argues that when the goal of teaching and learning is mastery, criterion-referenced measurements are essential, since only they are capable of indicating whether or not the mastery has been attained.

Given the assumption of mastery as a goal, this justification is logically unassailable. But should mastery be the goal? At first glance it is most attractive. Partial learning cannot possibly be as good as complete learning. Only a goal that is fully attained can be fully satisfying.

More than forty years ago Professor H. C. Morrison[3] at the University of Chicago developed and popularized a method of teaching based on the mastery of "adaptations" of understanding, appreciation or ability. These unlike skills, seemed to Professor Morrison not to be a matter of degree: "...the pupil has either attained it or he has not." To achieve such an adaptation the instructor should organize his materials into units, each focused on a particular adaptation. He should then follow a systematic teaching routine: teach, test, reteach, retest, to the point of actual mastery.

For a time Morrison's ideas were popular and influential. Around 1930, the Education Index listed 14 articles per year on applications of the system he had advocated. By 1950 the rate had fallen to about 5 articles per year. The Education Index volume for the 1967-68 academic year lists not a single article on this subject.

Recently the concept of mastery has been reintroduced into educational discussions as a corollary of various systems of individually prescribed instruction, and as a solution to the problem of individual differences in learning ability. Several authorities[4-8] have pointed out, quite correctly, that these differences can be expressed either in terms of how much a student can learn in a set time, or in terms of how long it takes him to learn a set amount. Why, they ask, should we not let time be the variable instead of amount learned?

Their arguments have great force when applied to basic intellectual skills that everyone needs to exercise almost flawlessly in order to live effectively in modern society. But these basic skills make up only a small fraction of what the schools teach and of what various people are interested in learning. Look about you at the various talents and interests that different people have developed. See how these differences complement each other in completing the diverse jobs that need doing in our society. Then ask why we should expect or require a student of a subject to

achieve the same level of mastery as every other student of that subject.

Ernest E. Bayless[9] made this point in his criticism of the Morrison method. He made another to which we have already alluded. Abilities, understandings and appreciations are, in the experience of almost everyone, not all-or-none adaptations. They are matters of degree. None but the simplest of them can ever be mastered completely by anyone. Hence any criterion of mastery is likely to be quite imperfect and arbitrary. To the extent that it is, our criterion-referenced measurements will also be imperfect and arbitrary as were the percent grades that norm-referenced measurements replaced fifty years ago.

To summarize, the major limitations of criterion-referenced measurements are these:

1. They do not tell us all we need to know about achievement.

2. They are difficult to obtain on any sound basis.

3. They are necessary for only a small fraction of important educational achievements.

Contrary to the impression that exists in some quarters, criterion-referenced measurements are not a recent development that modern technology has made possible and that effective education requires. The use of criterion-referenced measurements cannot be expected to improve significantly our evaluations of educational achievement.

It is true, of course, that norm-referenced measurements of educational achievement need to have content meaning as well as relative meaning. We need to understand not just that a student excells or is deficient, but what it is that he does well or poorly. But these meanings and understandings are seldom wholly absent when norm-referenced measures are used. We can make them more obviously present and useful if we choose to do so.

References

[1]Ebel, Robert L. "Content Standard Test Scores" Educational and Psychological Measurement.

[2]Starch, Daniel and Elliott, E. C. "Reliability of Grading High School Work in English," School Review, 20:442-57, 1912.

[3]Morrison, Henry C. The Practice of Teaching in the Secondary School, University of Chicago Press, 1926.

[4]Adkins, Dorothy C., Measurement in relation to the educational process. Educational and Psychological Measurement, 1958, 18, 221-240.

[5]Glasser, R. Instructional technology and the measurement of learning outcomes. American Psychologist, 1963, 18, 519-521.

[6]Carroll, John. A model for school learning. Teachers College Record, 1963, 64, 723-733.

[7]Gagne, R. The conditions of learning. New York: Holt, Rinehart, & Winston, 1965.

[8]Bloom, B. S. Learning for mastery. UCLA, CSEIP Evaluation comment, May 1968, 1 (2).

[9]Bayless, Ernest E. "Limitations of the Morrison Unit" Science Education 18:203-7 December 1934.

"CRITERION-REFERENCED TESTING IN THE CONTEXT OF INSTRUCTION[1]"

Anthony J. Nitko
Learning Research and Development Center
University of Pittsburgh

When we talk about criterion-referenced testing, we need to distinguish it from some traditional usages of the word criterion with which it tends to be confused. The term criterion has been used many times in psychometrics to refer to a second variable which we are interested in predicting. For example, an aptitude test is sometimes said to predict a criterion such as end of course grades or scores on an achievement test. Sometimes the validity of a test is described in terms of its correlation with some criterion (or criteria).

A second common usage of criterion has been that of criterion scores. The criterion score functions much the same as a cut-off score for some decision. In this context, expressions such as "working to criterion level" have been employed. For example, a statement like: "this student answered 50 percent of the test questions correctly, but has not reached the criterion level of performance which is answering 85 percent of the questions correctly."

Neither of these two usages of the term criterion is quite what is meant by criterion-referenced testing. It is useful, therefore, to review some of the background for criterion-referenced testing in order to more clearly describe it.

Criterion-Referenced Testing

Although it may be true that criterion-referenced tests were used earlier, the term can probably be attributed to Robert Glaser. It was first mentioned in connection with proficiency measurement in training (Glaser and Klaus, 1962) and later was applied to the measurement of educational achievement (Glaser, 1963). The motivation for this application to achievement measurement stemmed from a concern about the kind of achievement information required to make instructional decisions. Some instructional decisions concern individuals. For example, what kind of competence an individual needs in order for him to be successful in the next course in a sequence. Other decisions center around the adequacy of the instructional procedure itself. Tests which provided achievement information about an individual only in terms of how the individual compared with other members of the group, or which provided only sketchy information about the degree of competence he possessed with respect to some desired educational outcome, were not sufficient to make the kinds of decisions necessary for effective instructional design and guidance.

In his discussion, Glaser refers to two other people who had proposed similar ideas: John Flanagan (1951) and Robert Ebel (1962). Both the Flanagan and Ebel ideas, while similar to Glaser's, are different enough to warrant discussion.

The Flanagan reference is to his chapter on units, scores, and norms in Lindquist's (1951) Educational Measurement.

Flanagan distinguished between five types of descriptive information that are necessary in order to interpret broadly educational achievement data. In that discussion he made a distinction between a "standard of performance" and a "norm-performance." A standard of performance on a test is defined as a desirable model or a minimum goal we would like an individual to attain. A "norm-performance" is the present average performance or attainment with respect to a specific group or population. For example,

> The score of an individual as obtained on a French reading test might be at the tenth-grade norm. This gives little information about how well he reads various types of materials. The probable degree of comprehension of the individual in reading a typical French newspaper would provide a useful social standard for interpreting scores on a French reading test (pages 698-699).

He cautioned that it was unwise to use automatically and uncritically the present average test performance as the acceptable score for that test. The most fundamental piece of information that an achievement test should provide is a description of an individual's performance with respect to some defined body of content that can be interpreted without reference to the scores of other individuals or to norm groups.

Professor Ebel (1962) extended this distinction and presented two schemes for developing tests whose scores could be interpreted objectively and meaningfully without the use of norms. Of special emphasis are the content categories that the test items represent. One method would result in a display of selected test items along with descriptive information about how many of these items could be answered correctly by individuals at various total test score levels. For example, if 10 of the 50 mathematics items from the PSAT were displayed, it would be possible to make a statement such as: "Persons with a standard score of 500 on the mathematics section of the PSAT will, on the average, get 4 or 5 of these 10 items correct." The selected items are obtained by first sorting a large number of items into subject-matter content categories, such as: calculations with fractions, verbal problems, triangles, circles, and so on. Then the one item in each category that best discriminates between the high and low scoring groups on the entire test is selected to represent the content category. Data for assigning meaning to a score of 500

---

[1]Grateful acknowledgement is made to Robert Glaser and Richard L. Ferguson for their helpful comments on he draft manuscript.

The preparation of this paper was supported by the Learning Research and Development Center supported as a research and development center by funds from the United States Office of Education, Department of Health, Education and Welfare.

is obtained by finding how many of the ten items were answered correctly, on the average (the mode in this case), by those persons who had standard scores of 500. This is repeated for each standard score level.

A second, more basic, procedure for obtaining meaningful scores is to make the process by which the test is constructed systematic and explicit. This calls for a systematic sampling of test items, rather than a subjectively chosen collection of tasks. For, "unless the score is based on a systematic sample from a defined domain of tasks, it cannot provide a very sound basis for inferences as to the examinees' performance on similar collections of tasks (page 16)." As an illustration, tests were built that required the examinee to match definitions with words.

> "The tests were based on a spaced sample of 100 words from a specified dictionary. Explicit instructions were given (to the test constructors) for choosing a unique but representative sample, and for limiting the sample to words appropriate for the test. For each word the first synonym or defining phrase was copied from the dictionary ....These tests constitute one operational definition of the proportion of words in a certain dictionary for which a person 'knows' the meaning, and hence the size of his vocabulary in a certain sense (pages 24-25)."

The term "content-standard scores" was used to refer to the kind of scores derived from these tests. "Content" means that the score is based directly on the items comprising the test. "Standard" means both the common scale on which the scores are reported (percent in this case) and the fact that the process by which the test is constructed, administered, and scored is made explicit and objective. Thus, an individual's obtained score is referred directly to the domain of content for interpretation. This is contrasted to normative-standard scores which are interpreted by referring to the performance of other individuals. It should be noted that this is a different use of the word "standard" than was used by Flanagan, who used it in the sense of a minimum goal or a desired model.

In a way, Glaser (1962) combined both the notion of a desired model and the notion of a standard domain of content. He called for the specification of the type of behavior the individual is required to demonstrate with respect to the content. "The standard (or criterion) against which a student's performance is compared ... is the behavior which defines each point along the achievement continuum (page 519)." A criterion-referenced test, then, is one that is deliberately constructed to give scores that tell what kinds of behavior individuals with those scores can demonstrate (Glaser and Nitko, 1970).

As an illustration, consider the problem of assessing the competency of a student in elementary school geometry. Competency in elementary geometry can be analyzed into a number of behavior classes. A test can be constructed to measure these behaviors and to give scores that can be interpreted in terms of them. On such a test, a score of 30 might mean that, along with a number of lower level behaviors, the student is able to

> identify pictures of open continuous curves, lines, line segments, and rays; can state how these are related to each other; and can write symbolic names for specific illustrations of them. He can identify pictures of intersecting and non-intersecting lines and can name the point of intersection.

This score would also mean that the student could not demonstrate higher level behaviors such as

> identifying pictures that show angles; naming angles with three points; identifying the vertex of a triangle and an angle; identifying perpendicular lines; use a compass for bisection or drawing perpendiculars; and so on.

In like manner, a score of 20 might mean that the student could not demonstrate any of the behaviors implied by the higher scores, but could demonstrate all lower level behaviors, up to and including behaviors such as:

> naming the plane figures that comprise the faces of cubes, cones, pyramids, cylinders, and prisms; naming these solids; and identifying pictures of these solids.

It is apparent, then, that there are four characteristics inherent in criterion-referenced tests:

1. the classes of behaviors that define different achievement levels are specified as clearly as is possible before the test is constructed.

2. each behavior class is defined by a set of test situations (that is, test items or test tasks) in which the behaviors can be displayed in terms of all their important nuances.

3. given that the classes of behavior have been specified and that the test situations have been defined, a representative sampling plan is designed and used to select the test tasks that will appear on any form of the test.

4. the obtained score must be capable of expressing objectively and meaningfully the individual's performances characteristics in these classes of behavior.

Norm-Referenced Scores from Criterion-Referenced Tests

Norm-referenced testing is well known. When a test is constructed to yield scores that can be interpreted in such a way as to determine an examinee's relative location in a population or group of other examinees who took the same test, then we have a norm-referenced test. Scores derived for norm-referenced information are reported as percentiles, standard scores, grade-equivalents or age-equivalents. To obtain these scores, the mean, standard deviation, and sometimes the form of the distribution is pre-specified.

It should be obvious that criterion-referenced testing can yield norm-referenced information. Under certain circumstances both criterion-referenced information and norm-referenced information are needed to make a broad interpretation of an individual's test performance. Flanagan, Ebel, and Glaser all point this out.

In most circumstances one or the other kind of information is of primary concern. The test constructor can choose to maximize either criterion-referenced information or norm-referenced information, but seldom can he maximize both. Since norm-referenced scores derive most of their meaning from distributions in which we can distinguish one individual from another, judicious selection of test items with the help of statistical analysis will maximize this distinction. Such statistical selection of items for criterion-referenced tests makes little sense, however. The classes or domains of tasks which define a behavior are determined, insofar as is possible, before the test is constructed and then representative samples are drawn for inclusion on any test. To screen out some items for inclusion on a particular test because they possess desirable statistical characteristics will change the definitions of the behavioral categories (cf. Osburn, 1968). The kind of information desired when criterion-referenced tests are used is the behaviors an individual does or does not possess and whether or not the test yields meaningful normative-standard scores is often of secondary importance.

The Need for a Data Base

When one proceeds to build a criterion-referenced test he needs to be just as rigorous as when constructing a norm-referenced test. Given that the classes of behavior have been defined, empirical evidence is needed to support any contentions that the classes of test tasks do indeed reflect the behavior or competence of interest. There is a need for knowledge about test construction to become integrated with psychological knowledge and theory.

More often than not, a single verbal statement of a behavior implies that an individual ought to be able to perform quite a large domain of tasks. This is particularly true of instructional objectives, where generalization and transfer are of primary importance. These domains of tasks need to be systematically examined and, if necessary, stratified so that representative sampling can take place.

Most useful instructional objectives which are employed in curriculum design appear to be formulated as constructs. This is true because (1) the behavior that is referred to is most often stated in terms of a class of responses to a class of stimuli and (2) all of these statements are often tied together with psychological interpretations such as the need for prerequisites and the relationships among the objectives in the sequence of instruction. Specifications of the instructional objectives which are needed for criterion-referenced tests tend to avoid broad trait construct statements such as "reading ability." Thus, the job of building tests that have representative tasks defining classes of behavior becomes more difficult as the behaviors become more complex. It is easier to build tests to measure decoding skills than to measure reading comprehension. The basis for inference about "reading ability" for example, is observable performance on the specified domain of tasks into which reading ability can be analyzed, such as: reading certain types of passages aloud, identifying objects described in a test, rephrasing sentences in a certain way, carrying out written instructions, reacting emotionally to described events, and so on. It would seem, then, that criterion-referenced test builders need to

conduct many of the same kinds of construct validation studies as have been recommended for psychological tests and other kinds of achievement tests (Cronbach and Meehl, 1955; Cronbach, 1969).

Absolute Interpretation of Test Scores

Recently, Cronbach (1969) has called attention to the need for absolute interpretations of test performance. Criterion-referenced testing implies this also. Absolute interpretation refers to making judgments about a person's score in terms of what his performance on the test is and what that performance represents with respect to a defined domain of test tasks. It is contrasted with comparative or relative interpretations, by which judgments about a person's score are based on the scores of other individuals in the population or group to which he has membership. It is clear that the testing movement has given little attention to absolute interpretations (Cronbach, 1969).

Absolute interpretations can be extremely dangerous, however, if they are used inappropriately. Tests for which the domain of items is vaguely defined, for which the behaviors ellicted are indeterminate, and for which a representative sampling plan has been unspecified, are poor bases upon which to interpret scores in an absolute sense. Failure to perform proper analysis before test construction often leads to assessing only those educational goals that are easily measured. Such abuses are probably common in many classroom test interpretations -- and, perhaps, in much of what is currently passing for criterion-referenced testing! As Professor Ebel (1962; 1970) points out, such abuses are reminiscent of the criticisms of the percentage course grade and of objective testing early in this century.

These abuses, then, point more strongly toward the need for properly constructed criterion-referenced tests, based on well defined and instructionally meaningful behaviors, in situations where absolute interpretations tend to be made or where these interpretations need to be made. This means replacing much of the "art" of item writing with the technology of item writing; behavioral and task analysis, task construction, and domain specification. Such work is certainly not easy, but neither does it seem impossible. A few notable suggestions along these lines have been provided by Gagne (1969): Hively (1966); Hively,Patterson, and Page (1968); and Bormuth (1970).

Mastery

Criterion-referenced tests have been employed most often in instructional situations where the notion of mastery learning is advocated. One issue in which criterion-referenced testing has become entangled is that of determining mastery. Some propose that a cutoff or "criterion score" needs to be established and that each student must be taught until he obtains a score greater than or equal to this cutoff score. Some have argued that the cutoff score must be located at the upper extreme since flawless performance is desirable.

Nothing about criterion-referenced testing implies any of this. That criterion-referenced testing does not depend on a cutoff score has been mentioned previously. Further, criterion-referenced testing does not imply a value judgement about whether flawless performance is desirable. It only seeks to assess what the behavior is.

Whether using cutoff scores with tests is good or bad, is an empirical question although it is embedded in the ethical and decision network within which one operates. For example, given that certain terminal outcomes are desired and that an instructional sequence is specified, the question is: what level of performance is required at each point in the learning sequence in order to maximize success at the next point in the sequence and so on until the terminal learning is attained? This appears to be a transfer of learning problem and not one which is left entirely to subjective judgment. It is clear that such decisions cannot be based on poor information, such as a poorly constructed test, but must be based on the empirical findings of instructional psychology.

Related to criterion-referenced testing and mastery learning is the question of whether everyone needs to learn the same thing to the same degree and who imposes standards of competency. A reasonable discussion of this issue and its ethical implications is beyond the scope of the presentation. (For a cogent discussion of this issue in another context, see Bandura (1969). Much of that discussion seems to apply to instruction.) Nothing in the nature of criterion-referenced testing implies that anyone necessarily meet a given standard of competency, only that such levels of competency be defined in terms of performance.

A humanistic point of view would take into account the goals of the individual as related to the goals of society and allow the individual to participate in choosing and planning his learning experiences. If the individual desires to become a "master" and is motivated to achieve mastery, then of necessity we must provide him with the experiences which will facilitate his becoming a master and provide him with assessments so that he can evaluate his progress toward the goal he has chosen. To be sure, this point has been made by others. An interesting recent example of the successful application of behavioral analysis is that given by Zoellner (1969) with respect to the teaching of English composition. He states the problem in this way:

> "....the central failure of current compositional pedagogy....is its apparent inability to furnish the student-writer with anything but the most generalized specification for getting from one side of the writing situation (poor writing) to the other (good writing). What is urgently needed is a pedagogical technique which will supply the student-writer with a set of compositional specifications which are a) successively intermediate rather than ultimate, b) visible rather than invisible, c) uniquely adapted to the student's unique writing problem, and d) behavioral rather than historical, addressed to writing rather than the written word (page 274)."

The Need for Norm-Referenced Information

So far this discussion has emphasized criterion-referenced information. The need for norm-referenced information as well as criterion-referenced information should be apparent. It is useful under certain circumstances to know not only what level of competency an individual or group has or does not have, but also how that competency is related to other individuals or groups which are similar in composition, have similar educational experiences, or which have similar aspirations. It is also important to know relative standing in groups that are basically different.

But "useful" can only be interpreted in terms of purpose. In order to determine what kind of information to collect or to emphasize, one needs to know what kind of decision needs to be made. In some decision contexts norm-referenced information is inescapable. It has been pointed out that in some parts of the world it may be that it is financially impossible to offer advanced education to all individuals. Here relative competency and relative standing with respect to all such applicants for education becomes one of the most important types of information that is needed for decision-making. Whether such a stance is valid is beyond the scope of this presentation. The answer to such a question, however, will determine to a large extent the type of information the educational decision-maker will need and the kinds of observations and data that will have to be collected.

Criterion-Referenced Testing vs Norm-Referenced Testing

Is criterion-referenced information better than norm-referenced information? One cannot discuss the usefulness of one measurement procedure over another without knowing the context within which that information is needed and how it will be used. As Green (1969) has noted, considerations of measurement per se are wasteful in the overall decision-making process. Failing to consider the interrelationship between measurement and decision-making neglects the importance of deciding what additional data need to be collected before adequate decisions can be made.

There is a difference between taking measurement for scientific purposes and testing in instructional situations. The scientist is concerned with the identification and measurement of stable properties and variables. He seeks to determine general laws and rules for determining the relationships between these variables. He is dicipline-oriented and this dictates to a large extent the variables he chooses to measure and the way in which he measures them. In the practice of instruction one is concerned primarily about what each pupil desires to learn and how to maximize the learning he desires. What is learned is of primary importance and is usually defined in terms of acquired behavior and competence. Instruction provides the conditions by which this learning takes place. In a somewhat different context Lord (1968) speaks to this point.

> It should be clear that there are important differences between testing for instructional purposes and testing for measurement purposes. The virtue of an instructional test lies ultimately in its effectiveness in changing the examinee. At the end, we would like him to be able to answer every test item correctly. A measurement instrument, on the other hand, should not alter the trait being measured. Moreover, ..., measurement is most effective when the examinee knows the answers to only about half the test items. (page 2)

It is a platitudinous assertion that an educational system should provide for individual differences and should allow

students at every level of ability to develop and excel.. Several patterns of instructional procedures for adapting to individual differences as they appear in the school can be identified (Cronbach, 1967). One pattern occurs where educational goals and instructional methods are relatively fixed and inflexible. Individual differences are taken into account by dropping students along the way. The underlying rationale involved is that every child should "go as far as his abilities warrant." A second pattern of adaptation to individual differences is one in which the prospective future role of a student is determined and, depending upon this role, he is provided with an appropriate curriculum. For example, vocationally oriented students get one kind of mathematics and academically oriented students get a different kind of mathematics. Generally in this type of adaptation to individual differences the educational system has optional educational objectives, but within each option the instructional procedures are relatively fixed. A third pattern of adaptation to individual differences is one in which instructional procedures are varied to accommodate the differences in each student. Different students are taught differently, and the sequence of what is learned is not common to all students. One way in which this pattern is implemented is to provide a fixed mainstream instructional sequence and to branch students to remedial work when needed. Upon completion of remedial work the student is returned to the mainstream instruction. Another way of implementing this pattern is to begin with an assessment of a pupil's learning habits and attitudes, achievement and skills, cognitive style, etc. This information is used to guide the student through a course of instruction that is uniquely tailored to his goals. Thus, students would learn in different ways and attain different goals.

Each of these different patterns of instruction will require different kinds of measurement that result from different types of information requirements and instructional decision-making requirements. It is impossible, then, to speak of the strengths and weaknesses of criterion-referenced or norm-referenced testing in a vacuum. The merits of any testing program lie in the extent to which it provides useful information to the decision-maker, be he instructional designer, pupil, teacher, administrator, or the pupil at large.

Not only must this information be useful, but it must be usable as well. That is, the testing program must be designed into the instructional process so that the information that is required is easily obtained and available in a usable form at the time a decision needs to be made. Built into such an instructional system must be a procedure for constantly updating and redefining the adequacy of the decisions being made and the information upon which they are based.

When viewed in this way, the distinction between testing and instruction becomes less distinct, so that the learner can look toward testing for feedback concerning his accomplishments and for guidance toward his chosen goals.

References

Bandura, Albert Principles of behavior modification. New York: Holt, Rinehart, and Winston, 1969.

Bormuth, John On the theory of achievement test items. Chicago: University of Chicago Press, 1970.

Cronbach, Lee J. How can instruction be adapted to individual differences? In R. Gagne (ed.), Learning and individual differences. Columbus, Ohio: Charles E. Merrill Books, 1967, pp. 23-29.

Cronbach, Lee J. Validation of educational measures. In Proceedings of the 1969 Invitational Conference on Testing Problems. Princeton, New Jersey: Educational Testing Service, 1969.

Ebel, Robert L. Content-standard test scores. Educational and Psychological Measurement, 1962, 22, 15-25.

Ebel, Robert L. Some limitations of criterion-referenced measurement. Symposium address at the American Educational Research Association, Minneapolis, March, 1970.

Flanagan, John C. Units, scores, and norms. In E. F. Linquist (ed.), Educational Measurement. Washington, D.C.; American Council of Education, 1951, pp. 695-763.

Gagne, Robert M. Instructional variables and learning outcomes. In W. C. Wittrock and D. Wiley (eds.), Evaluation of instruction. New York: Holt, Rinehart, and Winston, 1969.

Glaser, Robert Instructional technology and the measurement of learning outcomes. American Psychologist, 1963, 18, 519-521.

Glaser, Robert and Klaus, David J. Proficiency measurement: Assessing human performance. In R. Gagne (ed.), Psychological principles in systems development. New York: Holt, Rinehart, and Winston, 1962, pp. 419-474.

Glaser, Robert and Nitko, Anothny J. Measurement in learning and instruction. In R. L. Thorndike (ed.), Educational Measurement. Washington, D.C.: American Council on Education, 1970. (in press)

Green, Burt F. Comments on tailored testing. In W. Holtzman (ed.), Computer-assisted instruction, testing and guidance. New York: Harper and Row, 1969.

Hively, Wells Preparation of a programmed course in algebra for secondary school teachers: A report to the National Science Foundation. Minnesota National Laboratory, Minnesota State Department of Education, 1966.

Hively, Wells, Patterson, H. L., and Page, S. A "universe-defined" system of arithmetic achievement tests. Journal of Educational Measurement, 1968, 5, 275-290.

Linquist, E. F. (ed.) Educational Measurement. Washington, D. C.: American Council on Education, 1951.

Lord, Fredrick M. Some test theory for tailored testing. Office of Naval Research Report. Princeton, New Jersey: Educational Testing Service, September, 1968.

Osburn, H. G. Item sampling for achievement testing. Educational and psychological measurement, 1968, 28, 95-104.

Zoellner, Robert Talk-write: A behavioral pedagogy for composition. College English, 1969, 30.

"CRITERION-REFERENCED TESTS"

Frederick B. Davis
University of Pennsylvania

A criterion-referenced test has been defined as "a measuring instrument deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards" (Glaser and Nitko, 1971). The interpretation of an examinee's score is wholly independent of the performance of other examinees in a "norm group" representative of some defined population. Ordinarily, scores are expressed as the number of items correct or the percentage of items correct.

At this point, consider whether a test properly constructed and scored in the manner described could be administered to samples of pupils representative of populations in which its use would be appropriate and whether percentile of populations in which its use would be appropriate and whether percentile ranks could be assigned to each raw score in each of the populations sampled. Obviously, this could be done and norm-referenced score interpretations could be made. Clearly, then, it is not the test itself that determines whether scores from it may be norm-referenced. Consequently, it might be wise to avoid describing tests as "criterion-referenced" or "norm-referenced." If we are to use these terms at all, they should be applied to scores, not to tests. The fact is that either type of score may be obtained for any test. Established principles of test theory indicate when either type is appropriate for a given test.

Although the term "norm-referenced scores" described reasonably well what it is intended to describe, there are persuasive reasons why the term "criterion-referenced scores" should be abandoned. First, the terms "criterion-referenced scores" and "norm-referenced scores" dichotomize all scores; hence, their use implies strongly that a test from which the former are derived has been carefully constructed to measure some defined criterion variable while a test from which the latter are derived has not been. In other words, educators and laymen are likely to infer that tests yielding criterion-referenced scores have higher "content validity" than tests yielding norm-referenced scores. This inference is categorically unjustified since any test can yield either type of score and since the content validity of a test is dependent mainly on the care and skill employed in designing and writing items for it and by the nature of the variable measured by it. Second, as Glaser and Nitko (1971) have pointed out, many people confuse criterion-referenced tests with tests yielding scores that have been correlated with an external criterion or with several such criteria in order to estimate the predictive validity coefficient or coefficients of such scores.

Among the terms that come to mind to replace "criterion-referenced scores" are "fixed-standard scores," "absolute scores," and "mastery-test scores." Of these, "fixed-standard scores" might be commonly confused with standard scores or normalized standard scores (like T-scores). The term "absolute scores" suggests that a true zero point has been established for the variable being measured, which is an unlikely accomplishment in educational measurement. "Mastery-test scores" is a phrase that grows out of the

historical development of instructional tests used informally in the classroom and coincides with what Glaser and Nitko appear to mean by criterion-referenced scores. They have stated that "the instructional process requires information about the details of the performance of the learner in order to know how instruction should proceed . . . When this performance has been attained by an individual learner to the degree required by the design of the instructional system, then the learner is said to have attained mastery of the instructional goal" (Glaser and Nitko, 1971). Therefore, it seems best to use the term "mastery-test scores" in place of "criterion-referenced scores."

Norm-referenced scores are used primarily to compare the performance of one examinee with that of others in a representative sample of some defined relevant population. They are less frequently used to differentiate among examinees in a sample; consequently, terms like "differentiation scores" or "differential scores" are not maximally appropriate. Instead, the phase "comparison scores" should be used in place of "norm-referenced scores."

Since time immemorial, teachers have, with varying degrees of success, measured the level of performance of their pupils on material or processes that have recently been taught by means of tests that meet Glaser and Nitko's definition of what the latter call criterion-referenced tests. In 1864, for example, Chadwick wrote that the Reverend George Fisher had prepared a book called the Scale Book, "which contains the numbers assigned to each degree of proficiency in the various subjects of examination . . . The numerical values for spelling . . . are made to depend upon the percentage of mistakes in writing from dictation sentences from works selected for the purpose, examples of which are contained in the Scale Book in order to preserve the same standard of difficulty" (Chadwick, 1864). By the 1920's individualized instruction theoretically gave every pupil the time and instruction needed to bring him to a predetermined level of accomplishment. This led to the development and use of diagnostic tests to guide instruction and of mastery tests to permit demonstration that certain prescribed skills and principles had been learned. The Winnetka Plan, the Morrison Unit-Mastery Plan, and the Dalton Plan made provision for frequent testing to make sure that pupils mastered the performance of specified skills or tasks at a predetermined level. In the Dalton Plan each pupil signed a contract to reach certain specified competencies in a given unit and was allowed to go on to the next unit only after he had demonstrated this level of competence on a mastery test.

Because instructional materials and accompanying diagnostic and mastery tests were not made generally available, these plans for individualizing instruction were abandoned in most schools. The majority of teachers simply lack the skill and the time required to formulate performance standards and to construct the hundreds of short diagnostic or mastery tests needed to guide individualized instruction in fairly large groups and to evaluate each pupil's performance with respect to these standards. Fortunately, as programmed courses of study became available during the 1950's that were made up of learning exercises revised experimentally to teach efficiently the competencies that constitute their behavioral objectives and subobjectives, short diagnostic and mastery tests were keyed to each step in the instructional process. These yield raw scores (usually number of items answered correctly) that are linked directly to performance standards determined in advance. Teaching, learning, and evaluation are woven together in such a way as to maximize the effectiveness of instruction for each individual pupil. Fears that these developments will stifle teacher initiative and professional development have been expressed. But these need not be justified. On the contrary, the teacher's role as a guide to individual learning activities, as a motivating agent, and as a classroom manager to engender an atmosphere conducive to learning can become more rewarding and more challenging than before.

Properly planned programs of evaluation should combine the frequent use of short diagnostic and mastery tests with the occasional use of standardized achievement tests, interest inventories, and specialized aptitude tests. Each type of test supplements the others. For what it may be worth, it is my opinion that many schools now use too few short diagnostic and mastery tests for instructional purposes and too many standardized tests. The reason for this is simply that most teachers do not have access to a supply of diagnostic and mastery tests keyed to the specific objectives of their instruction. I can see no practical solution to this problem short of creating and making available complete packages of behavioral objectives, instructional materials and procedures, and short diagnostic and mastery tests keyed to the objectives and prefiled in convenient, long-lasting cabinets. One part of this package without the others is nearly useless. Furthermore, as the instruction of Project PLAN has already shown, teachers must be tactfully and consistently guided in the use of such packages in their classrooms.

I should point out, however, that use of these packages for individualizing instruction and guiding learning will not prevent comparisons of the school achievement of different pupils. Say, for example, that the arithmetic curriculum in City A is organized for the first six years of schooling into carefully planned units of work leading to the attainment of 1,000 behavioral objectives. No pupil ever "fails" in arithmetic; every one spends as much time as he needs to attain each objective as it comes in the ordered sequence. At the end of two years a few pupils would have attained 400 or more objectives; others would have attained only 100 or fewer objectives. Parents are kept informed from time to time about the progress of their children in arithmetic by reports indicating, among other things, the number of objectives covered. If this information is not provided by the school officially, parents will compare notes and make estimates of their own. Naturally, they will ask teachers questions like, "Why has Sally Brown covered 200 objectives in arithmetic whereas my son has covered only 70 objectives in arithmetic? How many objectives should he have covered?" Inevitably, in one way or another differences in the number of objectives covered take on normative significance to parents and pupils alike.

The more instruction is individualized and made efficient, the more noticeable individual differences in rate and capacity for learning will become. Educators must accept this fact and deal with it. One solution would be that which some labor unions have adopted. A skilled man who works rapidly and efficiently is simply advised in one way or another to get back into line and conform to an acceptable display of ability. Another solution is to encourage diversity and the display of talent by providing a wide range of ways in which pupils can distinguish themselves and gain self-esteem.

This paper may perhaps best be concluded by discussing briefly the guidance that modern test theory can provide with respect to evaluation instruments like mastery tests. Specifically, what does test theory have to say about: 1) how to maximize the content validity of mastery tests; 2) how to make mastery-test scores legitimately interpretable in terms of specified performance standards; 3) how reliability coefficients and accuracy of measurement can be estimated for mastery-test scores; 4) how to evaluate the likelihood and seriousness of errors in determining whether a pupil has truly met predetermined standards of performance for any given instructional objective; 5) how long mastery tests need to be; and 6) what considerations influence the format of mastery-test items and how they should be scored.

First, the content validity of mastery-test scores can be maximized by conscientiously carrying out the conventional first step in the design of any achievement test. A detailed test outline must be prepared listing the specific objectives and subobjectives of the instructional unit to be evaluated. These must be expressed in terms of observable behaviors, to each of which one or more test exercises can be keyed. The display of substantive knowledge, skills and processes, attitudes, and feelings should be included, as required, in the populations of behaviors to be sampled by items.

Sampling the population of possible items for testing a specific objective may, in practice, be carried out by approximation procedures. For example, Glaser and Nitko (1971) mention the fact that the population of problems in the addition of 3, 4, and 5 addends with the restriction that each addend shall be a single-digit integer from 0 to 9 consists of 111,000 different problems. Proposals for rules to be followed in creating the desired number of items from a huge population have been discussed by several investigators.

In evaluating these proposals, item writers should recognize that the true tetrachoric intercorrelations of item scores (usually "1" or "0") of items drawn from the population of items covering any narrowly delimited objective will be close to unity. Therefore, minor deviations from a perfectly random sample of items are not likely to affect seriously a test's content validity (Wilks, 1938). It is important, however, for the test outline to specify the extent to which the direct effects of instruction and its transfer to analogous materials are to influence the test variance. For example, if a spelling rule is taught, its application to the words used in the instructional process is likely to be displayed better than its application to other words to which the rule also applies.

To make legitimate the interpretation of number-right scores, corrected raw scores, or percent-correct scores on any test, the content of the test must be homogeneous; that is, all of the items must measure the same variable (plus chance, of course). Such a test is said to be univocal. If a test is made up of a weighted composite of different skills, its raw scores do not properly represent successive levels of performance in any single objective. Consequently, when a pupil obtains less than a perfect score, the teacher cannot, on the basis of that score alone, determine what specific content or process he has not learned adequately. This situation and the uses to which mastery-test scores are put lead to the conclusion that such tests should be univocal. These considerations also indicate that a very large number of separate mastery tests are needed; thus, for practical

reasons they should be as short as possible. Since their reliability coefficients depend largely on their length, it is apparent that efficiency of measurement (that is, reliability per unit of time) is at a premium in such tests.

Whenever decisions are made wholly or partly on the basis of test scores, the frequency with which these decisions are in error becomes a matter of concern. This is partly because we want to be fair to the pupil and partly because errors lead to inefficiency in the instructional process. The errors can take two forms when we are using mastery-test scores to determine whether to advance a pupil to the next unit or to reteach the unit on which he has been tested: first, we can advance him when he should be held back; second, we can hold him back when he should be advanced. The incidence of such errors depends partly on the reliability coefficient of the determinations. Consider the reliability coefficient of scores on a dive-item test of skill in getting the main thought of five reading paragraphs administered to 421 college freshmen in 1940. Every examinee answered every item. The mean score was 2.97 items answered correctly; the variance of these scores was 1.21; the reliability coefficient was 0.18, and the standard error of measurement for any single score drawn at random from the 421 obtained was 1.00. Thus, an examinee who scored 3 points could easily have a true score anywhere between 2-4 points. The data show the caution with which only in separating the examinees into two groups: (1) those who obtained scores of 0-4, inclusive; and (2) those who obtained scores of 5 and are judged to have reached the predetermined level regarded as adequate for advancement to the next unit of instruction, the reliability coefficient for determining into which of the two groups each pupil belongs is 0.66, the cutoff score being 4.5. The procedure used to estimate this reliability coefficient for the "advance-no advance" determinations was recently provided by Livingston (1970). The result is in harmony with classical test theory. In general, the greater the difference between the cutoff score and the mean of the entire group, the more the reliability coefficient of the "advance-no advance" determinations (made by whole-number cutoff scores) vary with test length as predicted by the Spearman-Brown formula, we can estimate the number of items like those in the 5-item test that would be required to produce determinations of any desired reliability

If such determinations were the only basis for irrevocable placements of long-term importance to the pupils, we should insist on a reliability coefficient of the determination that would be above 0.90. But the penalty for misplacing a pupil at the end of a unit of instruction is not great because the decision can soon be changed by a teacher who observes his performance and each unit is likely to be short. Nevertheless, any errors of placement lower the overall efficiency of the instructional process so we want to hold their incidence to some acceptably low percentage, such as five our of every hundred decisions. Procedures for accomplishing this are well known. On the basis of the illustrative data that I have cited and other data of this kind that are available to me, I would hazard a guess that the majority of mastery tests would yield dichotomic classifications with acceptable accuracy if the tests were made up of 20-30 items.

If provisions can be made to score mastery tests by hand by qualified professional personnel (such as the classroom teachers themselves), the task of item writing is greatly simplified because a variety of item formats, including free-response questions, can be used. This freedom is especially helpful for making tests for use in the elementary school with children below the age of 11. Since examinees ordinarily have a chance to try every item in classroom tests, the conventional correction for chance success will not alter the rank order of number-right scores. However, when true-false items or multiple-choice questions with as few as 2-4 choices are used, corrected scores ordinarily provide considerably better estimates of the percent of the population of items sampled that is actually known by a pupil than are provided by number-right scores. It would be of interest to investigate the extent to which partial knowledge and misinformation balance each other in the conventional correction formula when it is used with mastery tests of the type we have been discussing. Very little information is available about this matter and analytic formulations are not helpful.

In conclusion, it seems safe to say that mastery and diagnostic tests supplement standardized survey tests in educational evaluation. Each type serves an important educational need better than other types. Educators, therefore, are not faced with the problem of choosing between them but should concentrate their efforts in using all evaluation instruments to maximum advantage as the need for each of them appears.

References

Chadwick, E. Statistics of educational results. The Museum, Quarterly Magazine of Education, Literature, and Science, 1864, 3, 480-4.

Glaser, R. and Nitko, A. J. Measurement in learning and instruction. In Thorndike, R. L., ed. Educational Measurement. Washington: American Council on Education, 1971.

Levingston, S. A. The Reliability of Criterion-referenced Measures. Baltimore: Center for the Study of Social Organization of the Schools, The Johns Hopkins University, Report No. 73, July 1970.

Wilks, S. S. Weighting Systems for linear functions of correlated variables when there is no dependent variable Psychometrika, 1938, 8, 23-40.

- - -